ROBUST INFERENCE VIA Lq-LIKELIHOOD

by

Yichen Qin

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2013

Abstract

Robust inference is of great importance in modern statistics. In this dissertation, we introduce a series of robust statistical procedures based on the concept of Lq-likelihood [1]. The Lq-likelihood function partially preserves the desired properties of the log-likelihood function. Moreover, it provides remarkable robustness, on which we can develop robust statistical procedures. The tuning parameter q of the Lq-likelihood makes our robust statistical procedures more flexible; because when $q \to 1$, the Lq-likelihood reduces to the traditional log-likelihood. Therefore, we can use q to adjust the efficiency-robustness trade off as well as the bias-variance trade off. In this dissertation, we first introduce a new robust estimator called maximum Lq-likelihood estimate (MLqE) and derive its properties from a robust statistics point of view. We also develop a robust testing procedure — the Lq-likelihood ratio test (LqLR) — and demonstrate its effectiveness on contaminated data. We further move to the problem of robust estimation of mixture models and propose an expectation maximization algorithm for Lq-likelihood (EM-Lq). Finally, we develop a robust clustering technique and provide an application of our technique to brain graph data.

ABSTRACT

Advisor: Dr. Carey E. Priebe

Primary Reader: Dr. Carey E. Priebe

Secondary Reader: Dr. Bruno Jedynak

Acknowledgments

I would like to express my deepest gratitude to my advisor Professor Carey E. Priebe for his excellent guidance, patience and encouragement throughout the entire process. I feel very fortunate to be able to work under Professor Priebe. His invaluable suggestions, optimism and cheerful spirit always provide me strength to overcome difficulties. He has given me the perfect combination of academic supervision and freedom, which ignites my great interest in robust statistics. Professor Priebe has made the PhD program fun, rewarding, and life-changing for me. It is because of Professor Priebe that I decided to pursue an academic career. I owe him a great deal of gratitude.

I also would like to thank Professor Daniel Q. Naiman and Professor Bruno Jedynak who provided me constant help throughout the entire PhD program. Professor Naiman was my academic advisor when I entered the PhD program and supported me in the early stages of my PhD program. Professor Jedynak has also been extremely supportive of my research and served on my candidacy exam and graduate board oral exam. Meanwhile, I have worked with Professor Naiman and Professor Jedynak on

various research projects and have learned a great deal from them. Moreover, both of them provided essential help during my job search process. I want to thank them sincerely for their constant support.

There are many people at Johns Hopkins University who deserve my thanks. I want to thank Professor J Tilak Ratnanather for giving me research directions, guiding me through the application part of my dissertation research, and serving on my graduate board oral exam. I am grateful to Professor Donald Geman who supervised my research during my second year as a PhD student. Special thanks go to Professor Joshua Vogelstein who provided data and helpful comments for my research which eventually become one chapter in my dissertation. Last but not least, I would also like to thank Professor Yingyao Hu in the economics department for his help in serving as the chair of my graduate board oral exam committee.

Finally, I would like to thank my parents for their unconditional love and support and my wife for her love, understanding and faith in me.

Dedication

This thesis is dedicated to my parents and my wife.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Robust Statistics

Robust statistics has been of great importance in modern statistics. There has been extensive research on the treatment of robust statistics. We understand that statistics is a subject dealing with collecting, analyzing and interpreting data. These procedures are well studied under a collection of strict assumptions. However, when these assumptions are not satisfied, traditional statistical methods fail to maintain effectiveness. On the other hand, robust statistics studies the effect of assumption violation and proposes remedies. Robust statistical methods perform about as well as traditional methods under the assumptions. Meanwhile, robust methods also maintain good performance when assumptions are violated, whereas traditional methods break down.

Robust statistics has been well studied since [2]. Several important concepts have been introduced in the context of robust statistics, for example, influence function [3] and breakdown point [4]. In this dissertation, we will introduce another class of estimator which is a special case of the M-estimator. We will study its properties in terms of estimation and testing.

1.2 Preliminaries

1.2.1 Exponential Family

Commonly used distributions, for example, Normal, Binomial, Poisson, Gamma, Beta, etc., all belong to one important class of distribution — exponential family — whose properties have been well studied [5]. A distribution family $\{f_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, is called a p-parameter exponential family if there exits real valued functions $\eta_1(\theta), ..., \eta_p(\theta)$, $B(\theta)$, $T_1(x), ..., T_p(x)$ and $h(x)$ (with $x \in \mathbb{R}^d$) such that the density function of f_θ can be written as

$$f(x;\theta) = h(x) \exp\Big[\sum_{j=1}^{p} \eta_j(\theta) T_j(x) - B(\theta)\Big], \quad x \in \mathcal{X} \subset \mathbb{R}^d$$

where the vector $(T_1(x), ..., T_p(x))$ is the sufficient statistic for the parameter of such a distribution.

For example, for a normal distribution, it is clear that

$$\varphi(x; u, \sigma^2) = \exp\left[\frac{u}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2}(\frac{u^2}{\sigma^2} + \log(2\pi\sigma^2))\right]$$

which means the natural parameters of a normal distribution are

$$\eta_1(u, \sigma^2) = \frac{u}{\sigma^2},$$
$$\eta_2(u, \sigma^2) = -\frac{1}{2\sigma^2}.$$

1.2.2 Kullback-Leibler Divergence

The Kullback-Leibler divergence (KL) is an asymmetric "distance" between two distributions. It measures the difference of two distributions: g and h. It is defined as

$$KL(g||h) = \int g(x) \log \frac{g(x)}{h(x)} dx.$$

Notice that $KL(g||h) \geq 0$.

1.2.3 MLE for Misspecified Models

Sometimes the assumed model may provide an incorrect description of the data. For example, the true data generating process $g(\cdot)$ does not belong to the assumed

class of distributions $\{f : f \in \mathcal{F}\}$. This is called model misspecification. To some extent, all models are misspecified. The consequences of using a misspecified model are of particular concern in many disciplines. In this section, we briefly introduce the properties of MLE for misspecified models. In [6], the author has shown that MLE is a consistent (and asymptotically normally distributed) estimator of f^*, which is the model in the assumed class of models that minimizes the KL distance between the true data generating process g and the assumed model $\mathcal{F}$, i.e.,

$$f^* = \arg\min_{f \in \mathcal{F}} KL(g||f).$$

Therefore f^*, the element in $\mathcal{F}$ that has the least KL distance from g, can be considered as a projection of g onto $\mathcal{F}$. For more detailed explanation, please refer to [6].

1.2.4 Mixture Models

A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. It is defined as

$$f(x;\theta) = \sum_{j=1}^{k} \pi_j f_j(x;\theta_j), \quad \sum_{j=1}^{k} \pi_j = 1 \quad \pi_j = 1 > 0$$

where π_j is the component weight, $f_j(x;\theta_j)$ is the component density with component parameter θ_j, and k is the complexity of the mixture model (i.e., number of compo-

nents). A mixture model is at the heart of many statistical problems such as image segmentation and clustering. Among all mixture models, a Gaussian mixture model (GMM) is the most frequently used. The Gaussian mixture model is one type of mixture model with all component densities being normal distributions. The parameter of such a model is $\theta = (\pi_1, ..., \pi_{j-1}, u_1, ..., u_j, \sigma_1^2, ..., \sigma_j^2) \in \Theta$, where

$$\Theta = (0,1)^{k-1} \times (-\infty, \infty)^k \times (0, \infty)^k \subset \mathbb{R}^{3k-1}.$$

The likelihood under the mixture model is unbounded at the boundary of the parameter space Θ. The MLE of θ of such a mixture model as a global maximizer of the likelihood function does not exist. However, the largest local maximizer of the likelihood function has been proved to be a consistent estimator of the parameter. To obtain such an estimate, we apply the expectation maximization algorithm (EM), which is an iterative algorithm that usually converges to the local maximizer. In this chapter, we properly set the initial values of the iterative procedure to obtain estimates so that we avoid the singularities of likelihood surface. A detailed explanation of mixture models and MLE for mixture models can found in [7].

1.3 Maximum Likelihood Estimation

A likelihood function $L(\mathbf{x}, \theta)$ represents the likelihood of observing the current sample under the assumed model with parameter θ. Hence, the likelihood is a function

of θ.

The likelihood function is one of the most important concept in statistics. Since it was first used by R.A. Fisher in 1922 [8] under the context of "method of maximum likelihood," there has been a great amount of research based on Fisher's original idea.

The likelihood function is a cornerstone for frequentist statistics as well as Bayesian statistics. Therefore, the study of the likelihood function is critical to the development of statistics.

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model.

1.4 Lq-Likelihood

Ferrari and Yang [1] proposed an alternative of likelihood function called Lq-likelihood. Instead of using the log function, [1] introduces another function called the Lq function,

$$L_q(u) = \frac{u^{1-q} - 1}{1 - q},$$

where $q > 0$. Notice that when $q \to 1$, $L_q(u) \to \log(u)$.

Accordingly, given a sample $\mathbf{x} = (x_1, ..., x_n)$, the L$q$-likelihood is defined as

$$L_q(\mathbf{x}, \theta) = \sum_{i=1}^{n} L_q(f(x_i; \theta)).$$

The newly introduced Lq-likelihood appears to be more robust than the traditional log-likelihood when $q < 1$. This is because the log function is unbounded from below, i.e., $\log u \to \infty$ as $u \to 0$. On the other hand, the L_q function is bounded when $q < 1$, i.e., $L_q(u) \to -1/(1-q)$ as $u \to 0$. Figure 1.1 compares two functions. For $q < 1$, the L_q function is above the log function. For $q > 1$, the L_q function is below the log function. Both cases preserve concavity. When $q \to 0$, L_q tends to $x - 1$.

To see the effect on the Lq/log-likelihood surface, we present a simple example. Suppose we have a sample $\mathbf{x} = (x_1, ..., x_n)$, $n = 10$, from a normal distribution $N(u, \sigma^2 = 1)$ with known variance $\sigma^2 = 1$. The Lq/log-likelihood surface against parameter u is plotted in Figure 1.2. As we can see from the figure, depending on $q < 1$ or $q > 1$, the Lq-likelihood is uniformly above or below the log-likelihood, which is consistent with the property of the L_q function shown in Figure 1.1. That is,

$$L_q(u) > \log(u) \text{ for } 0 < q < 1,$$

$$L_q(u) < \log(u) \text{ for } q > 1.$$

However, consider changing one of the observations, say, x_{10} to a much larger value, say, $x_{10}^{\text{new}} = x_{10} + 5$. This change can be thought of as a measurement error or an

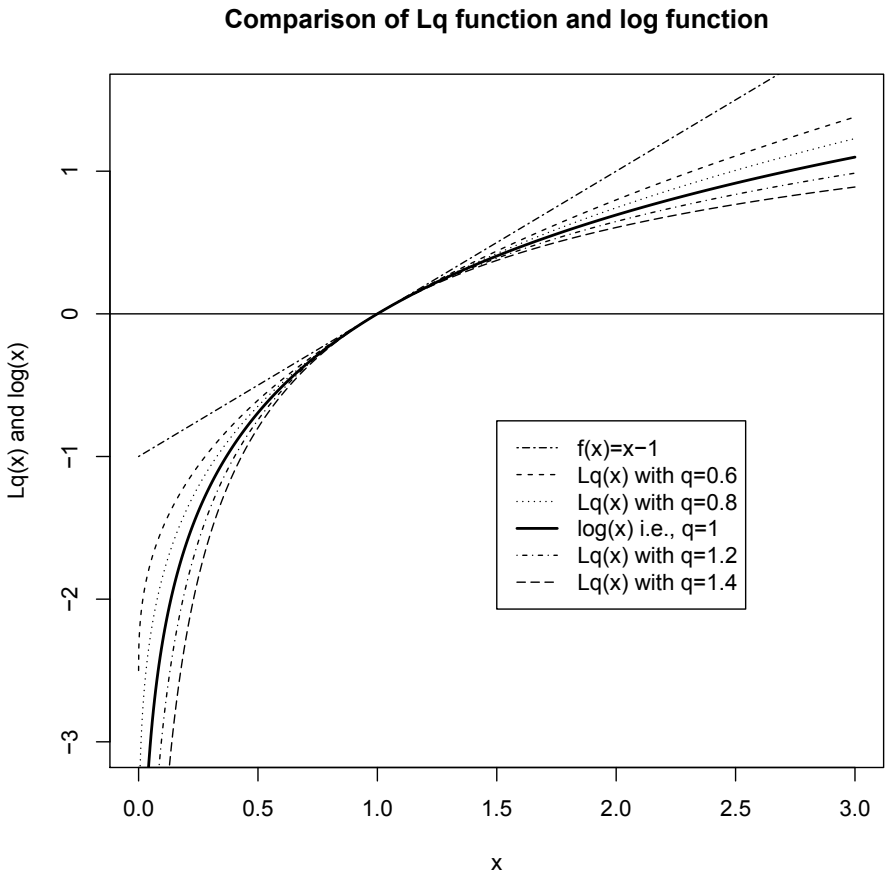


Figure 1.1: Comparison of the Lq and log functions.

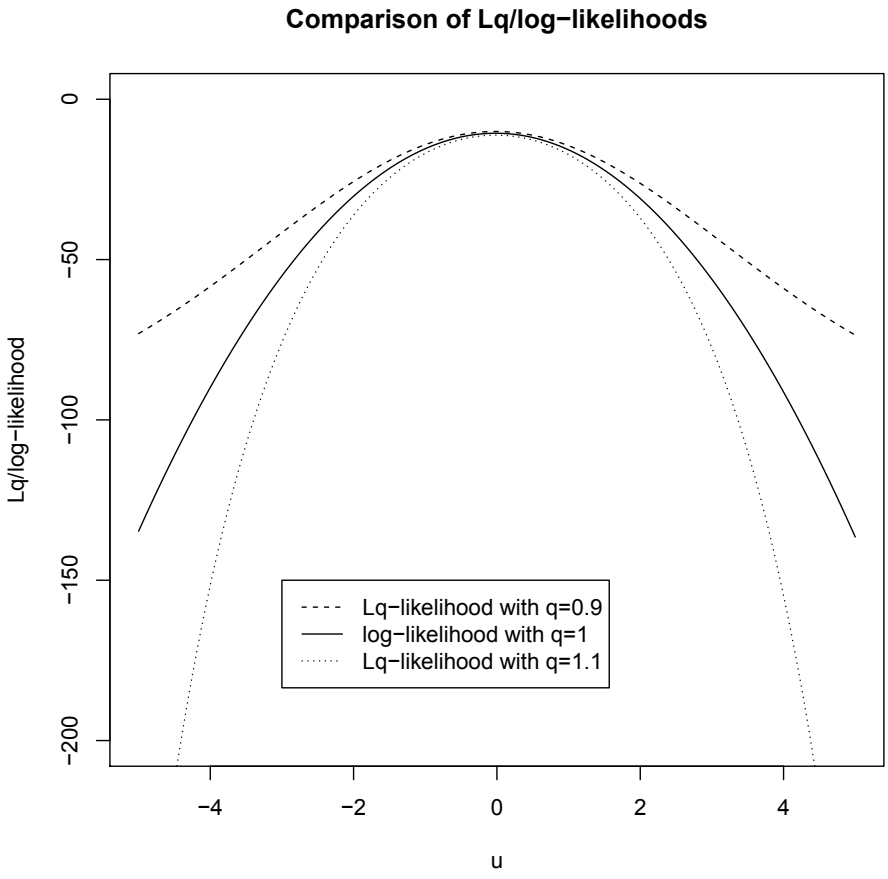


Figure 1.2: Comparison of the Lq/log likelihoods against parameter u for a sample from $N(u, \sigma^2 = 1)$.

outlier. We plot the new Lq/log-likelihood surfaces based on the modified sample along with the old Lq/log-likelihood surfaces in Figure 1.3. We can see that the log-likelihood surface is greatly changed, whereas Lq-likelihood ($q < 1$) surface is much less sensitive to the perturbation of the data. On the other hand, the Lq-likelihood ($q > 1$) surface is much more sensitive to the perturbation than the log-likelihood surface. From this figure, we can have an idea of how to make use of the robustness of the Lq-likelihood function. Throughout this dissertation, we focus mainly on $q < 1$; however, the case of $q > 1$ is also of great importance in other research areas.

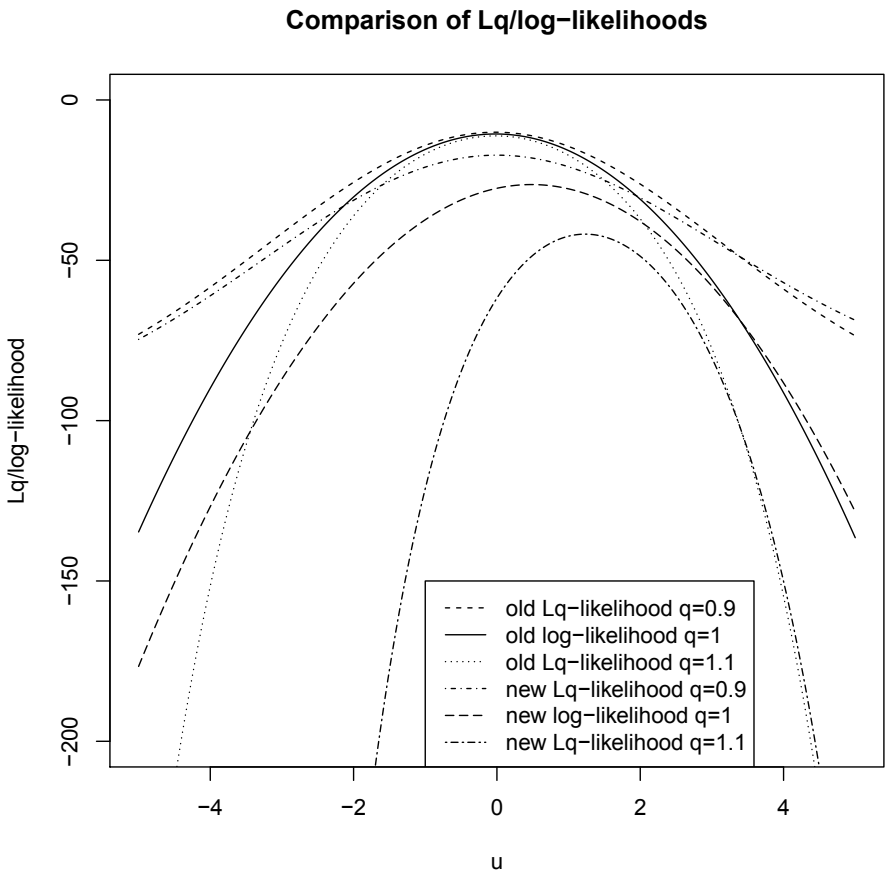


Figure 1.3: Comparison of the Lq/log likelihoods against parameter u. Old likelihoods are for the original sample. New likelihoods are for the modified sample with an outlier.

Chapter 2

Maximum Lq-Likelihood Estimation

In this chapter, we introduce maximum Lq-likelihood estimation and study its properties.

2.1 Definitions and Basic Properties

First, let us start with the traditional maximum likelihood estimation. Suppose data X follows a distribution with probability density function f_θ parameterized by $\theta \in \Theta \subset \mathbb{R}^p$. Given the observed data $\mathbf{x} = (x_1, ..., x_n)$, the maximum likelihood

estimate is defined as

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta}\{\sum_{i=1}^{n}\log f(x_i;\theta)\}.$$

Similarly, the maximum Lq-likelihood estimate [1] is defined as

$$\hat{\theta}_{\mathrm{ML}q\mathrm{E}} = \arg\max_{\theta\in\Theta}\sum_{i=1}^{n} L_q(f(x_i;\theta)),$$

where $L_q(u) = (u^{1-q}-1)/(1-q)$ and $q > 0$. By L'Hopital's rule, when $q \to 1$, $L_q(u) \to \log(u)$. The tuning parameter q is called the distortion parameter, which governs how distorted L_q is away from the log function. Based on this property, we conclude that the MLqE is a generalization of the MLE.

Define

$$U(x;\theta) = \nabla_\theta \log f(x;\theta) = \frac{f'_\theta(x;\theta)}{f(x;\theta)},$$
$$U^*(x;\theta,q) = \nabla_\theta L_q(f(x;\theta)) = U(x;\theta)f(x;\theta)^{1-q}.$$

We know that $\hat{\theta}_{\mathrm{MLE}}$ is a solution of the likelihood equation $0 = \sum_{i=1}^{n} U(x_i;\theta)$. Similarly, $\hat{\theta}_{\mathrm{ML}q\mathrm{E}}$ is a solution of the Lq-likelihood equation

$$0 = \sum_{i=1}^{n} U^*(x_i;\theta,q) = \sum_{i=1}^{n} U(x_i;\theta)f(x_i;\theta)^{1-q}. \tag{2.1}$$

It is easy to see that $\hat{\theta}_{\mathrm{ML}q\mathrm{E}}$ is a solution to a weighted version of the likelihood equation that $\hat{\theta}_{\mathrm{MLE}}$ solves. The weights are proportional to the power transformation of the probability density function, $f(x_i;\theta)^{1-q}$. When $q < 1$, the MLqE puts more weight on the data points with high likelihoods, and less weight on the data points with low likelihoods. The tuning parameter q adjusts how aggressively the MLqE distorts the weight allocation. The MLE can be considered as a special case of the MLqE with equal weights.

In particular, when f is a normal distribution, our $\hat{\mu}_{\mathrm{ML}q\mathrm{E}}$ and $\hat{\sigma^2}_{\mathrm{ML}q\mathrm{E}}$ satisfy

$$\hat{\mu}_{\mathrm{ML}q\mathrm{E}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_i, \tag{2.2}$$

$$\hat{\sigma^2}_{\mathrm{ML}q\mathrm{E}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i (x_i - \hat{\mu}_{\mathrm{ML}q\mathrm{E}})^2, \tag{2.3}$$

where $w_i = \varphi(x_i; \hat{\mu}_{\mathrm{ML}q\mathrm{E}}, \hat{\sigma^2}_{\mathrm{ML}q\mathrm{E}})^{1-q}$ and φ is a normal probability density function.

From equations (2.2) and (2.3), we conclude that the MLqE of the mean and the variance of a normal distribution are just the weighted mean and weighted variance. When $q < 1$, the MLqE gives smaller weights for data points lying in the tail of the normal distribution, and puts more weights on data points near the center. By doing so, the MLqE becomes less sensitive to outliers than the MLE at the cost of introducing bias into the estimation. A simple and fast re-weighting algorithm is available for solving (2.2) and (2.3) [1]. Details of the algorithm are described in Chapter 7.

2.2 Consistency and Bias-Variance Trade Off

Before discussing the consistency of the MLqE, let us look at the MLE first. It is well studied that the MLE is quite generally a consistent estimator. Suppose the true distribution $f_0 \in \mathcal{F}$, where $\mathcal{F}$ is a family of distributions; we know that

$$f_0 = \arg\max_{g \in \mathcal{F}} E_{f_0} \log g(X),$$

which shows the consistency of the MLE. However, when we replace the log function with the L_q function, we do not have the same property.

We first define $f^{(r)}$, a transformed distribution of f called the escort distribution, as

$$f^{(r)} = \frac{f(x;\theta)^r}{\int f(x;\theta)^r dx}. \tag{2.4}$$

We also define $\mathcal{F}$ to be a family of distributions that is closed under such a transformation (i.e., $\forall f \in \mathcal{F}$, $f^{(r)} \in \mathcal{F}$). Equipped with these definitions, we have the following property:

$$f_0^{(1/q)} = \arg\max_{g \in \mathcal{F}} E_{f_0} L_q(g(X)).$$

Thus we see that the maximizer of the expectation of the Lq-likelihood is the escort distribution ($r = 1/q$) of the true density f_0. In order to also achieve consistency for

the MLqE, [1] lets q tend to 1 as n approaches infinity.

For a parametric distribution family $\mathcal{G} = \{f(x;\theta) : \theta \in \Theta\}$, suppose it is closed under the escort transformation (i.e., $\forall\theta \in \Theta$, $\exists\theta' \in \Theta$, s.t. $f(x;\theta') = f(x;\theta)^{(1/q)}$). We have a similar property, $\tilde{\theta} = \arg\max_{\theta\in\Theta} E_{\theta_0} L_q(f(X;\theta))$, where $\tilde{\theta}$ satisfies $f(x;\tilde{\theta}) = f(x;\theta_0)^{(1/q)}$.

We now understand that, when maximizing the Lq-likelihood, we are essentially finding the escort distribution of the true density, not the true density itself, so our MLqE is asymptotically biased. However, this bias can be compensated by variance reduction if the distortion parameter q is properly selected. Take the MLqE for the normal distribution for example. With an appropriate $q < 1$, the MLqE will partially ignore the data points on the tails while focusing more on fitting data points around the center. The MLqE obtained this way is possibly biased (especially for the scale parameter), but will be less volatile to a significant change of data on the tails, hence, a good example of bias-variance trade off. The distortion parameter q can be considered as a tuning parameter that adjusts the magnitude of the bias-variance trade off.

2.3 Confidence Intervals

There are generally two ways to construct confidence intervals for the MLqE. One is parametric, the other is nonparametric. In this section, we discuss the univariate

case. The multivariate case can be obtained via a natural extension.

For the parametric approach, we know that the MLqE is an M-estimator, whose asymptotic variance is available. In order to have the asymptotic variance be valid, we need the sample size to be reasonably large so that the Central Limit Theorem applies. However, in our application, the MLqE deals with small or moderate sample sizes in most cases. So the parametric approach is not ideal, but it does provide a guideline to evaluate the estimator.

The second approach is the nonparametric bootstrap method. We create bootstrap samples from the original sample, and calculate their MLqEs for all bootstrap samples. We further calculate the lower and upper quantiles of these MLqEs, and call these quantiles the lower and upper bounds of the confidence interval. This method is model agnostic, and works well with the MLqE.

2.4 MLqE for Exponential Family

A family of distribution is called an exponential family if there exist functions $\eta(\theta)$, $B(\theta)$, $T(x)$ and $h(x)$, such that the density of the distribution family can be written as

$$f(x;\theta) = h(x) \exp \eta(\theta)T(x) - B(\theta).$$

Suppose $\mathbf{x} = (x_1, ..., x_n)$ follows $f(x;\theta)$ where f belongs to an exponential family

with natural parameter θ. We know that MLE of θ is consistent, that is

$$\hat{\theta}_{\mathrm{MLE}} \xrightarrow{p} \theta, n \to \infty.$$

However, we do not have such a property for MLqE. Instead, we have

$$\hat{\theta}_{\mathrm{ML}q\mathrm{E}} \xrightarrow{p} \frac{\theta}{q}, n \to \infty.$$

For example, if $f(x)$ is a normal distribution $\varphi(x; u, \sigma^2)$, then the natural parameters of such a distribution family are $\frac{u}{\sigma^2}$ and $-\frac{1}{2\sigma^2}$, therefore, the MLqE for the mean and variance are

$$\hat{u}_{\mathrm{ML}q\mathrm{E}} \xrightarrow{p} u, n \to \infty$$
$$\hat{\sigma^2}_{\mathrm{ML}q\mathrm{E}} \xrightarrow{p} \frac{\sigma^2}{q}, n \to \infty.$$

Chapter 3

Robust Hypothesis Testing via Lq-Likelihood

In this chapter, we introduce a robust testing procedure — the Lq-likelihood ratio test (LqLR)— and show that, for the special case of testing the location parameter of a symmetric distribution in the presence of gross error contamination, our test dominates the Wilcoxon-Mann-Whitney test at all levels of contamination.

3.1 Introduction

The likelihood ratio test (LR) is one of the most frequently used statistical tools in many areas of scientific research. However, only under a collection of strict assumptions does the LR obtain its assumed optimal performance. It is known that

its performance degrades significantly due to a merely mild violation of model assumptions. In an attempt to overcome this problem, we propose a robust testing procedure — the Lq-likelihood ratio test (LqLR) — using the newly developed concept of Lq-likelihood [1]. Under a gross error model, the performance of the LqLR is compared favorably to the LR and other nonparametric tests, such as the Wilcoxon-Mann-Whitney test [9, 10] and the sign test [11]. In the special case of testing the location parameter of a symmetric distribution, our testing procedure uniformly beats the Wilcoxon test and the sign test at all levels of contamination.

Our study of the LqLR focuses on the context of a gross error model $h(x) = (1-\epsilon)f(x;\theta) + \epsilon g(x)$, where f is our "idealized" model with the parameter θ that we are interested in testing, g is the measurement error component (or the contamination component), ϵ is the contamination ratio. With $\epsilon > 0$, h represents the true data generating process which is a small deviation from the "idealized" model f. For a data set generated by h, the majority of the data points (i.e., roughly a proportion of $1-\epsilon$) come from f, whereas the rest of the data points (from g) are usually considered measurement errors or outliers.

The measurement error problem has been one of the most practical problems in Statistics. Suppose we have some measurements $X = (X_1, X_2, ..., X_n)$ generated by a scientific experiment. X follows a distribution f_θ with an interpretable parameter θ, our parameter of interest. However, we do not observe X, rather, we observe $X^* = (X_1^*, X_2^*, ..., X_n^*)$ where most of the $X_i^* = X_i$, but there are a few outliers due to

human errors, or instrument malfunction, or simply an errorful observation process. In other words, X^* is X contaminated with gross errors. Under such circumstances, using data X^*, we still have θ as the target parameter for our hypothesis testing or estimation [5]. To overcome this problem, we introduce the LqLR.

Robust statistics has been well studied for the past 50 years. It addresses the problem of model assumption violation and proposes remedies for this issue. A robust statistical procedure performs nearly optimally when model assumptions are valid and still maintains good performance when the assumptions are violated. A robust procedure should be able to produce a valid conclusion regardless of a few bad or contaminated data points. Within the subject of robust statistics, there is relatively less research on testing than estimation [3, 4]. This is partially because the setting for hypothesis testing is more complex than estimation.

[2] suggested a form of likelihood ratio as $T(\mathbf{x}) = \prod_{i=1}^{n} \max(c', \min(c'', p_1(x_i)/p_0(x_i)))$. The tuning parameters c' and c'' are brought into the equation to address the effect of outliers whose likelihood is exceedingly small and causes the ratio $p_1(x_i)/p_0(x_i)$ to approach zero or infinity. However, hard thresholding using c' and c'' not only causes problems for maximization or minimization, it also induces sensitivity to the thresholds. On the other hand, the LqLR can be considered as a smooth version of the Huberized likelihood ratio test.

The structure of our chapter is as follows. We begin with a brief introduction of Lq-likelihood, and compare a "light" version of our LqLR with the log-likelihood based

test statistic in terms of relative efficiency in Section 3.2. We further introduce our major contribution — the LqLR — in Section 3.3 and prove its robustness properties via the analysis of the asymptotic distribution. We also discuss several related issues such as identifying the critical values. Numerical results of our test are presented in Section 3.4. We discuss the selection of the tuning parameter q in Section 3.5 and demonstrate the superior performance of our test compared to the LR, the Wilcoxon test, and the sign test. We provide discussion and conclusions in Section 3.6 and relegate the proofs to Appendix (Chapter 7).

3.2 Lq-Likelihood Based Test Statistic

3.2.1 Lq-Likelihood

A likelihood function measures the likelihood of the observed sample $\mathbf{x} = (x_1, ..., x_n)$ under the hypothesized model. It is defined as $L(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$, where f is the hypothesized model with $\theta \in \Theta \subset \mathbb{R}^d$. Usually it is more convenient to work with the log-likelihood, $l(x; \theta) = \log L(\mathbf{x}; \theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$. [1] introduced the L$q$-likelihood which is defined as $\sum_{i=1}^{n} L_q(f(x_i; \theta))$. It essentially replaces the log function by the Lq function with a tuning parameter $q > 0$. The Lq function is defined as $L_q(u) = (u^{1-q} - 1)/(1 - q)$ for $q \neq 1$, and $L_q(u) = \log u$ for $q = 1$. Notice that when $q \to 1$, $L_q(u) \to \log u$. Throughout this chapter, we assume $0 < q \leq 1$.

To estimate θ based on $\mathbf{x}$, maximum likelihood estimation is usually used: $\hat{\theta}_{\text{MLE}} =$

$\arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(x_i;\theta)$. Alternatively, we can use maximum Lq-likelihood estimation (MLqE), $\hat{\theta}_{\mathrm{ML}q\mathrm{E}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} L_q(f(x_i;\theta))$. For ML$q$E, we solve the L$q$-likelihood equation, $0 = \sum[f'_\theta(x_i)/f_\theta(x_i)]f_\theta(x_i)^{1-q}$, which is a weighted version of the likelihood equation. When $q < 1$, data points with high (or low) likelihoods are assigned large (or small) weights. As $q \to 1$, the MLqE becomes MLE.

The reason we gain robustness from the Lq-likelihood is that the Lq function is bounded from below for $0 < q < 1$. It is easily seen that $L_q(u) \geq -1/(1-q)$, whereas $\log(x) \to -\infty$ when $x \to 0^+$. In this case, if we have an outlier, say x_1, which gives a very small value of $f(x_1;\theta)$, then $\sum \log f(x_i;\theta)$ approaches $-\infty$, no matter whether θ gives high likelihood for x_2, ... , x_n, i.e., large values of $f(x_2;\theta)$, ... , $f(x_n;\theta)$. On the other hand, since $L_q(u)$ is bounded, it limits the effect of one particular data point on the quantity $\sum L_q(f(x_i;\theta))$. Therefore, the Lq-likelihood surface is much more stable than the log-likelihood surface against a perturbation of a small portion of the data.

3.2.2 MLqE as the Test Statistic and its Relative Efficiency

To show the advantage of Lq-likelihood in terms of relative efficiency, we temporarily use the MLq estimate of θ ($\hat{\theta}_{\mathrm{ML}q\mathrm{E}}$ with $0 < q < 1$) as our test statistic until we introduce our Lq-likelihood ratio test (LqLR) in Section 3.3. We denote this light version of our test statistic as $T_{q,n}$. We compare $T_{q,n}$ ($q < 1$) with $T_{1,n}$ ($q = 1$), which

is the ML estimate $\hat{\theta}_{\text{MLE}}$.

Suppose we are given the observed data $\mathbf{x} = (x_1, ..., x_n)$ from a pdf $f(x;\theta)$. We want to test the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$ with a size of α.

Definition 3.2.1. Define $u_q(\theta) = E_\theta[T_{q,n}]$, $\psi_q(x;\theta) = \frac{\partial}{\partial\theta} L_q(f(x;\theta)) = \frac{f'_\theta(x;\theta)}{f(x;\theta)} f(x;\theta)^{1-q}$, and $\psi'_q(x;\theta) = \frac{\partial^2}{\partial\theta^2} L_q(f(x;\theta)) = \frac{\partial}{\partial\theta}\psi_q(x;\theta)$. When $q = 1$, we have $\psi_1(x;\theta) = f'_\theta/f$.

Theorem 3.2.1. The asymptotic distribution of $T_{q,n}$ is $\sqrt{n}(T_{q,n} - u_q(\theta)) \sim N(0, V_q(\theta))$, where $V_q(\theta) = E[\psi_q(X;\theta)^2]/E[\psi'_q(X;\theta)]^2$. When $q = 1$, we have $E[\psi_1(X;\theta)^2] = -E[\psi'_1(X;\theta)]$. Hence, $\sqrt{n}(T_{1,n} - u_1(\theta)) \sim N(0, 1/E[\psi_1(X;\theta)^2])$ which attains the Cramér-Rao lower bound.

Proof. The proof follows from the asymptotic normality of the M-estimator. □

We use $T_{q,n}$ as our test statistic and reject H_0 when $T_{q,n}$ is large. To maintain the size of α, we reject H_0 when $\frac{T_{q,n} - u_q(\theta_0)}{\sqrt{V_q(\theta_0)/n}} \geq C_{q,n}$. Notice that $C_{q,n} \to z_{1-\alpha}$ when $n \to \infty$, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution.

It is straightforward to prove that $T_{q,n}$ with $0 < q \leq 1$ satisfies the assumptions 1 - 4 of [12] pp 371-372, which are restated in Appendix (Chapter 7). Therefore, we have

Definition 3.2.2. The efficacy of $T_{q,n}$ $(0 < q \leq 1)$ is $c_q = \frac{u'_q(\theta_0)}{\sqrt{V_q(\theta_0)}}$, where u'_q is the derivative of u_q.

Theorem 3.2.2. For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 + \frac{\delta}{\sqrt{n}}$, the limit of the

power of $T_{q,n}$ $(0 < q \leq 1)$ is $\Pi_{q,n} \to \Phi(c_q\delta - z_{1-\alpha})$, where $\Phi(\cdot)$ is the cumulative distribution function for standard normal distribution.

Proof. See [12], pp 372, Theorem 11. □

Now we study how the relative efficiency between $T_{1,n}$ and $T_{q,n}$ $(0 < q < 1)$ changes as the level of contamination increases. Suppose data follows a gross error model $h(x;\theta,\epsilon) = (1-\epsilon)f(x;\theta) + \epsilon g(x)$, where $f(x,\theta)$ is the idealized model, θ is the location parameter that we want to test for, g is the contamination component, and ϵ is the contamination ratio. Notice that when $\epsilon = 0$, we have $h = f$.

In this case, the expectation of $T_{q,n}$ under h becomes $u_q(\theta) = E_h T_{q,n}$. The asymptotic distribution of $T_{q,n}$ (Theorem 3.2.1) is still valid with V_q redefined as $V_q(\theta) = \frac{E_h[\psi_q(X;\theta)^2]}{E_h[\psi_q'(X;\theta)]^2}$. The null hypothesis $H_0 : \theta = \theta_0$ is tested against the alternative hypothesis $H_1 : \theta > \theta_0$. From [12], the relative efficiency between $T_{1,n}$ and $T_{q,n}$ is defined as $e_{q,1} = (c_q/c_1)^2 = \left(\frac{u_q'(\theta_0)/\sqrt{V_q(\theta_0)}}{u_1'(\theta_0)/\sqrt{V_1(\theta_0)}}\right)^2 = \frac{V_1(\theta_0)}{V_q(\theta_0)}\left(\frac{u_q'(\theta_0)}{u_1'(\theta_0)}\right)^2$.

Theorem 3.2.3. Suppose f and g are distributions that are symmetric about θ. The relative efficiency between $T_{1,n}$ and $T_{q,n}$ is $e_{q,1} = \frac{V_1(\theta_0)}{V_q(\theta_0)}$. The limiting power of $T_{q,n}$ becomes $\Pi_{q,n} \to \Phi(\frac{\delta}{V_q(\theta_0)} - u_\alpha)$.

Proof. When f and g are distributions that are symmetric about θ, it is easy to see that $u_1(\theta) = u_q(\theta) = \theta$; therefore, $u_1'(\theta) = u_q'(\theta) = 1$. Applying the result on the definition of relative efficiency proves the theorem. □

For a concrete example, let us assume h, the true data generating process, to be a

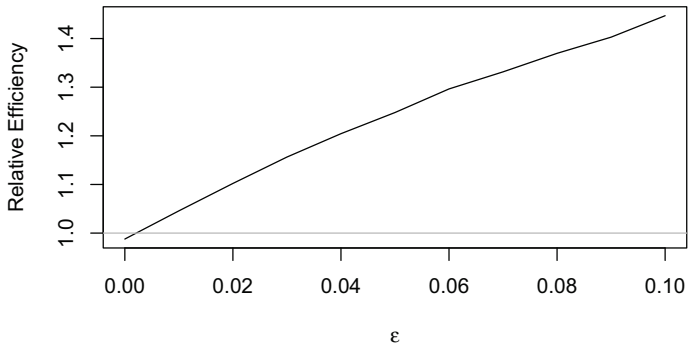


Figure 3.1: Relative efficiency $e_{q,1}$ as a function of contamination ratio ϵ.

gross error model $h(x;\theta,\epsilon) = (1-\epsilon)\varphi(x;\theta,1) + \epsilon\varphi(x;\theta,10)$ where the normal distribution $\varphi(x;\theta,1)$ corresponds to f and the normal distribution $\varphi(x;\theta,10)$ corresponds to g. By setting $q = 0.9$, we plot $e_{q,1}$ as a function of ϵ in Figure 3.1. As we can see from the figure, $e_{q,1}$ starts below 1 and gradually increases above 1. This implies that, in order to achieve the same level of power, it takes $T_{1,n}$ fewer data points than $T_{q,n}$ when there is no contamination. On the other hand, when contamination level gradually increases, it takes $T_{1,n}$ more data points to get the same power of $T_{q,n}$. Note that the ratio is only slightly below 1 when contamination ratio is 0, but significantly higher than 1 when contamination ratio increases over 1%. Hence, we have successfully traded efficiency at zero-contamination for robustness at heavy contamination.

3.3 Lq-Likelihood Ratio Test

3.3.1 Lq-likelihood Ratio Test Statistic

With the success of the previous section, we can continue to define a Lq-likelihood ratio test. Before we state the definition, let us briefly review the traditional likelihood ratio test (LR).

Suppose we have data $\mathbf{x} = (x_1, ..., x_n)$. The null and alternative hypotheses are given by $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$ where Θ_0 is the null parameter space and Θ_1 is the alternative parameter space. The LR test statistic is $\Lambda(\mathbf{x}) = \sup_{\theta\in\Theta_0} L(\mathbf{x}, \theta) / \sup_{\theta\in\Theta_0\cup\Theta_1} L(\mathbf{x}, \theta) = L(\mathbf{x}, \hat{\theta}_0)/L(\mathbf{x}, \hat{\theta}_1)$, where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the ML estimates of θ within Θ_0 and $\Theta_0 \cup \Theta_1$, respectively. Normally we use the equivalent test statistic $D(\mathbf{x}) = -2\log\Lambda(\mathbf{x}) = -2\sum_{i=1}^{n} \log f(x_i, \hat{\theta}_0) + 2\sum_{i=1}^{n} \log f(x_i, \hat{\theta}_1) \geq 0$. We reject the null hypothesis when we have a large value of $D(\mathbf{x})$. Naturally, we can define the Lq-likelihood ratio test (LqLR) as

$$D_q(\mathbf{x}) = -2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_{q,0})) + 2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_{q,1}))$$

where $\hat{\theta}_{q,0}$ and $\hat{\theta}_{q,1}$ are MLq estimates of θ within the parameter spaces Θ_0 and $\Theta_0\cup\Theta_1$, respectively. We reject the null hypothesis when we have a large $D_q(\mathbf{x})$. Note that when $q = 1$, the LqLR becomes the LR.

3.3.2 Asymptotic Distribution

In this section, we derive the asymptotic distribution of the LqLR test statistic. For simplicity, we assume a simple null hypothesis $H_0 : \theta = \theta_0$, a composite alternative hypothesis $H_1 : \theta \neq \theta_0, \theta \in \Theta$ and a 1-dimensional parameter space $\Theta \subset \mathbb{R}$. Hence, $D_q(\mathbf{x}) = -2\sum_{i=1}^{n} L_q(f(x_i, \theta_0)) + 2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_q))$, where $\hat{\theta}_q$ is the MLq estimate of θ. For such a test statistic, we have

Theorem 3.3.1. The asymptotic null distribution of $D_q(\mathbf{x})$ is given by

$$D_q(\mathbf{x})\Big| H_0 \xrightarrow{d} \frac{E[\psi_q(X;\theta_0)^2]}{-E[\psi_q'(X;\theta_0)]}\chi_1^2,$$

where χ_1^2 is a random variable following a Chi-square distribution with a degree of freedom 1.

Proof. We apply the Taylor expansion on the first term of $D_q(\mathbf{x})$ at $\theta = \hat{\theta}_q$ and obtain

$$\begin{aligned}
D_q(\mathbf{x}) =& -2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_q)) - 2(\theta_0 - \hat{\theta}_q)\sum_{i=1}^{n} \frac{\partial}{\partial\theta} L_q(f(x_i, \hat{\theta}_q)) \\
&- (\theta_0 - \hat{\theta}_q)^2 \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} L_q(f(x_i, \tilde{\theta})) + 2\sum_{i=1}^{n} L_q(f(x_i, \hat{\theta}_q)) \\
=& -(\theta_0 - \hat{\theta}_q)^2 \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} L_q(f(x_i, \tilde{\theta})) \\
=& -\Big(\frac{\sqrt{n}(\theta_0 - \hat{\theta}_q)}{\sqrt{V_q(\theta_0)}}\Big)^2 V_q(\theta_0)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} L_q(f(x_i, \tilde{\theta})),
\end{aligned}$$

where $\tilde{\theta}$ is a point between θ_0 and $\hat{\theta}_q$. We understand that $\frac{\sqrt{n}(\theta_0 - \hat{\theta}_q)}{\sqrt{V_q(\theta_0)}}\Big| H_0 \xrightarrow{d} N(0,1)$,

therefore, $\Big(\frac{\sqrt{n}(\theta_0-\hat{\theta}_q)}{\sqrt{V_q(\theta_0)}}\Big)^2\Big|H_0 \xrightarrow{d} \chi_1^2$. We know $\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}L_q(f(x_i,\tilde{\theta}))\Big|H_0 \xrightarrow{P} E[\psi_q'(X;\theta_0)]$ and $V_q(\theta_0) = \frac{E[\psi_q(X;\theta_0)^2]}{E[\psi_q'(X;\theta_0)]^2}$. So Slutsky's Theorem completes the proof. □

Notice that when $q = 1$, $E[\psi_1(X;\theta_0)^2] = -E[\psi_1'(X;\theta_0)]$, so D_1 follows a regular Chi-square distribution, which is the LR case. When $q < 1$, D_q follows a "distorted" Chi-square distribution with the distortion captured by the ratio $\frac{E[\psi_q(X;\theta_0)^2]}{-E[\psi_q'(X;\theta_0)]}$.

When we have contamination in the data (i.e. data are generated by a gross error model $h = (1-\epsilon)f + \epsilon g$), the results in Theorem 3.3.1 are still valid but the expectation is taken under h. Now let us discuss the asymptotic distribution of D_q under contamination. First, we make the following definitions for the sake of simplicity in notion:

Definition 3.3.1. Define

$$A(\epsilon,q) = E_h[\psi_q(X;\theta)^2] = E_h\Big[\Big(\frac{f_\theta'}{f}f^{1-q}\Big)^2\Big]$$
$$B(\epsilon,q) = -E_h[\psi_q'(X;\theta)] = E_h\Big[-\frac{f_\theta''}{f}f^{1-q} + q\Big(\frac{f_\theta'}{f}\Big)^2 f^{1-q}\Big]$$

Clearly, $A(\epsilon=0,q=1) = B(\epsilon=0,q=1)$.

Theorem 3.3.2. Based on Definition 3.3.1, the asymptotic distribution of $D_q(\mathbf{x})$ under the gross error model h is given by

$$D_q(\mathbf{x})\Big|H_0 \xrightarrow{d} \frac{A(\epsilon,q)}{B(\epsilon,q)}\chi_1^2.$$

When $q = 1$, D_1 becomes the LR test statistic. When $\epsilon = 0$ and $q = 1$, we have $D_1(\mathbf{x})\Big|H_0 \xrightarrow{d} \chi^2_1$ which is the case of the LR test using data with no contamination. On the other hand, when $\epsilon > 0$ (i.e., data with contamination) and $q = 1$, we have that $D_1(\mathbf{x})$ also follows a distorted Chi-square distribution with the "distortion" captured by the following theorem.

Theorem 3.3.3. Suppose f, g and $h = (1-\epsilon)f+\epsilon g$ are three symmetric distributions with the same mean θ. Assume that f satisfies the regularity conditions for the maximum likelihood estimation. Assume that g has a relatively fat tail compared to f, in the sense that $E_g[f''_\theta(X,\theta)/f(X,\theta)] > 0$. Then it holds that $A(\epsilon, q = 1) > B(\epsilon, q = 1) > 0$ for $\epsilon > 0$, or equivalently,

$$\frac{A(\epsilon, q = 1)}{B(\epsilon, q = 1)} > 1, \text{for } \epsilon > 0. \tag{3.1}$$

When f is a normal distribution, the condition $E_g[f''_\theta/f] > 0$ becomes $\sigma^2_g > \sigma^2_f$ where σ^2_g and σ^2_f are the variances of g and f.

Proof. See Appendix (Chapter 7) for proof. □

Remarks: For the condition in Theorem 3.3.3 ($0 < E_g[f''_\theta/f] = \int f''_\theta \cdot g/f dx$), please note that when $g = f$, we have $0 = E_f[f''_\theta/f] = \int f''_\theta dx$. Therefore, this condition means that the ratio g/f inflates the quantity $\int f''_\theta dx$ to be positive. When g has a fat tail distribution compared to f, then g/f is greater than 1 when $|x|$ is large and g/f is less than 1 when $|x|$ is small. Meanwhile, f is a distribution satisfying the

regularity conditions of the maximum likelihood estimation and usually has a bell shape. Therefore, f''_θ takes positive values at large $|x|$ and negative values at small $|x|$.

Theorem 3.3.3 implies that the LR test statistic D_1 with contaminated data follows an "inflated" Chi-square distribution under the null hypothesis. The same phenomenon is present for the asymptotic distribution under the alternative hypothesis (i.e., an "inflated" non-central Chi-square distribution). As the inflation of the asymptotic distribution becomes more serious, the null and alternative distributions become flatter, therefore, the overlap between the null distribution of $D_1|H_0$ and the alternative distribution of $D_1|H_1$ will become larger (see Figure 3.4 in Section 3.3.3 for more details). This explains the degradation of the power when contamination is brought into the data. In order to control the degradation of the power, we need to control the inflation of the asymptotic distribution. The following theorems illustrate how we can control of the inflation of the asymptotic distribution of D_q with $0 < q < 1$.

As ϵ increases away 0, we have the following theorem.

Theorem 3.3.4. Under the same assumptions as in Theorem 3.3.3, it holds that

$$\frac{\partial}{\partial \epsilon}|A(\epsilon, q=1) - B(\epsilon, q=1)| > 0 \text{ for } \epsilon \geq 0.$$

Proof. See Appendix (Chapter 7) for proof. □

Theorem 3.3.4 implies that as ϵ increases away 0, the discrepancy between $A(\epsilon, q =$

1) and $B(\epsilon, q = 1)$ also increases. That is to say, as we have more contamination in the data, the asymptotic distribution of D_1 (the LR test statistic) becomes more inflated.

However, by setting $q < 1$, we have the following theorem.

Theorem 3.3.5. Under the same assumptions as in Theorem 3.3.3 and an additional assumption stated in Appendix (Chapter 7), we have

$$\frac{\partial}{\partial q}\frac{\partial}{\partial \epsilon}|A(\epsilon, q) - B(\epsilon, q)| > 0 \text{ for } \epsilon \geq 0, q \leq 1. \tag{3.2}$$

Proof. See Appendix (Chapter 7) for proof. □

Theorem 3.3.5 implies that by setting $q < 1$, we can alleviate the inflation of the ratio A/B as a function of ϵ. With the help of Theorem 3.3.3, Theorem 3.3.4 and Theorem 3.3.5, we can further demonstrate

Theorem 3.3.6. With the same conditions as in Theorem 3.3.5, for any $\epsilon > 0$, there exists a C such that when $C < q < 1$, we have

$$\Big|\frac{A(\epsilon, 1)}{B(\epsilon, 1)} - 1\Big| > \Big|\frac{A(\epsilon, q)}{A(\epsilon, q)} - 1\Big|. \tag{3.3}$$

Proof. See Appendix (Chapter 7) for proof. □

What Theorem 3.3.6 means is that by setting $q < 1$, we can pull the ratio $A(\epsilon, q)/B(\epsilon, q)$ towards 1. The effect of $q < 1$ on the ratio $A(\epsilon, q)/B(\epsilon, q)$ can be

used to offset the inflation effect of contamination $\epsilon > 0$ on the ratio $A(\epsilon, q)/B(\epsilon, q)$. Therefore, by setting $q < 1$ we alleviate the magnitude of inflation of the asymptotic distribution under the null and alternative hypotheses and hence create protection for the power of the test.

In summary, we have proved that the divergence between $A(\epsilon, q)$ and $B(\epsilon, q)$ are much more serious for $q = 1$ than for $q < 1$. Even though we have $A(\epsilon = 0, q = 1) = B(\epsilon = 0, q = 1)$ at zero contamination, the loss of power at $\epsilon > 0$ due to the divergence between $A(\epsilon > 0, q = 1)$ and $B(\epsilon > 0, q = 1)$ is not affordable for any likelihood-based statistical tests. On the other hand, by setting $q < 1$ we lose the exact equality at zero contamination, that is, $A(\epsilon = 0, q < 1) \neq B(\epsilon = 0, q < 1)$, but the divergence between A and B is much less, and hence the power is greatly preserved. We want to point out that, by setting $q < 1$, we trade the exact equality of $A = B$ at $\epsilon = 0$ for much less divergence between A and B at heavy contamination $\epsilon > 0$. In the following section, we will illustrate our findings through numerical examples.

3.3.3 Simulation Study on Asymptotic Distribution

In this section, we study the asymptotic distribution under the normal distribution assumption. Let us assume f is a normal distribution with unknown mean θ

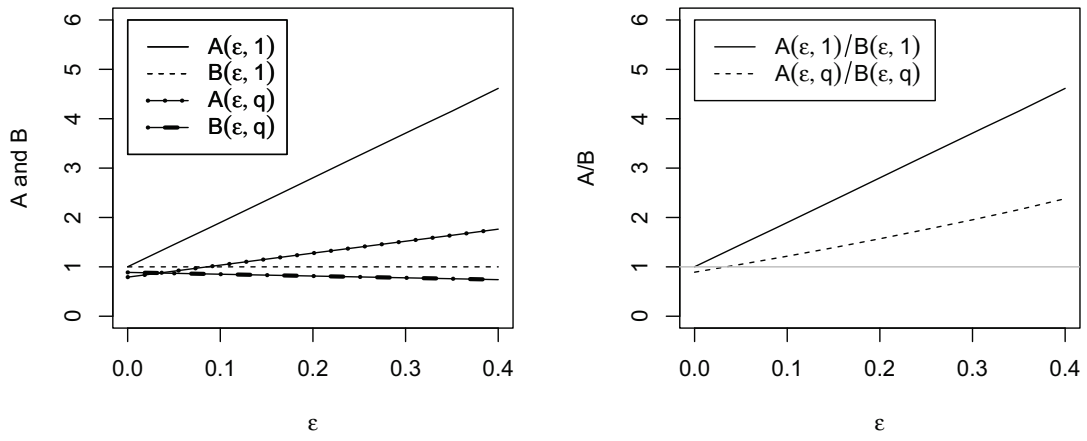


Figure 3.2: Left panel: Comparison of $A(\epsilon, q)$, $B(\epsilon, q)$, $A(\epsilon, 1)$ and $B(\epsilon, 1)$ at different levels of contamination. Right panel: Comparison of $A(\epsilon, q)/B(\epsilon, q)$ and $A(\epsilon, 1)/B(\epsilon, 1)$ at different levels of contamination.

and known variance σ_f^2. For $0 < q \leq 1$, we have $\psi_q(x, \theta) = \frac{x-\theta}{\sigma_f^2}\varphi(x; \theta, \sigma_f^2)^{1-q}$ and $\psi_q'(x, \theta) = \left[(1-q)\frac{(x-\theta)^2}{\sigma_f^4} - \frac{1}{\sigma_f^2}\right]\varphi(x; \theta, \sigma_f^2)^{1-q}$. We present a simulation study of $A(\epsilon, q)$ and $B(\epsilon, q)$ in Figure 3.2 for $q = 1$ and $q = 0.95$. We calculate these two quantities under the gross error model $h(x) = (1-\epsilon)\varphi(x; 0, 1) + \epsilon\varphi(x; 0, 10)$ as functions of ϵ. In the left panel of Figure 3.2, we see that as the contamination becomes greater, the difference between $A(\epsilon, 1)$ and $B(\epsilon, 1)$ increases faster than the difference between $A(\epsilon, q)$ and $B(\epsilon, q)$ for $q < 1$. In the right panel, we plot the ratio $A(\epsilon, q)/B(\epsilon, q)$ and $A(\epsilon, 1)/B(\epsilon, 1)$. We see that the ratio $A(\epsilon, 1)/B(\epsilon, 1)$ diverges from 1 as contamination increases, whereas the ratio $A(\epsilon, q)/B(\epsilon, q)$ for $q = 0.95$ is closer to 1 as contamination increases.

We further plot the ratio $A(\epsilon, q)/B(\epsilon, q)$ as a function of ϵ and q in a contour plot

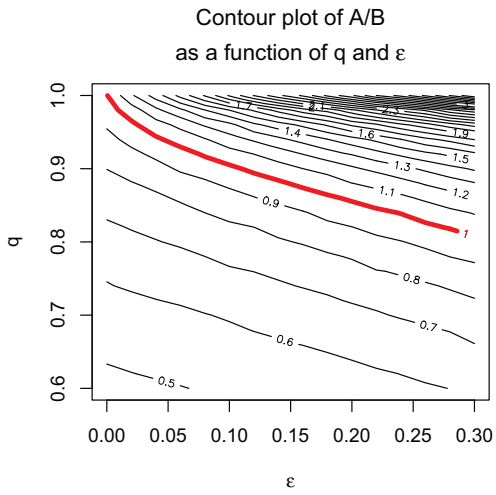


Figure 3.3: Contour plot of $A(\epsilon, q)/B(\epsilon, q)$ as a function of ϵ and q. As we can see, by setting $q < 1$, we can always decrease the ratio A/B and pull it back to 1.

in Figure 3.3. We highlight the level of 1 in bold red curve (i.e. not inflated). As we can see, when we stand at $q = 1$, the ratio A/B increases as ϵ increases. However, by decreasing q below 1, we can always find a value of q such that the ratio A/B is closer to 1. The red curve indicates the optimal q at different levels of contamination ϵ.

In Figure 3.4, we provide a simulation study of the asymptotic distributions under the null and alternative hypotheses. We are testing the mean of a normal distribution with known variance. We simulate data (sample size $n = 1000$) from $h(x) = (1 - \epsilon)\varphi(x; \theta, 1) + \epsilon\varphi(x; \theta, 50)$. We set $\theta = 0$ and simulate the distribution of the test statistic D_q under the null hypothesis. We set $\theta = 0.19$ and calculate the distribution of the test statistic under the alternative hypothesis. We change the contamination coefficient from 0 to 0.4, and set $q = 1, 0.97, 0.6$ to compare the effect of contamination

on these distributions. In the first row of Figure 3.4, we have $q = 1$ (i.e., LR). As the contamination increases, both the distributions of D_1 under the null and alternative hypotheses become flatter (i.e., inflated), which results in power degradation. In the second row, we have $q = 0.97$. We see that instead of having the inflated Chi-square distribution, the distributions under the null and alternative hypotheses are less affected by the contamination. This is because the ratio is pulled back to 1 by setting $q < 1$. In the third row, we have $q = 0.6$, which provides much more protection. The distributions are much less affected, and they hardly change as the contamination increases. However, it is worth noting that, in the lower left figure ($q = 0.6$ and $\epsilon = 0$), the null and alternative distributions overlap more than they do in the upper left figure ($q = 1$ and $\epsilon = 0$), which means that by setting $q < 1$ we lose power of the test at zero contamination. This figure illustrates how we gain robustness using the Lq-likelihood with $q < 1$ and trade for robustness by giving up a little power at zero contamination.

3.3.4 Bootstrap Estimation of the Critical Value

In the previous section, we discussed the variation of the null distribution of D_q at different levels of contamination. From our research we find that that the null distribution depends on the magnitude of contamination. However, in practice, we hardly know the contamination ratio ϵ and other properties of the contamination component g (i.e., variance σ_g^2 and etc.), therefore, we do not know the exact null

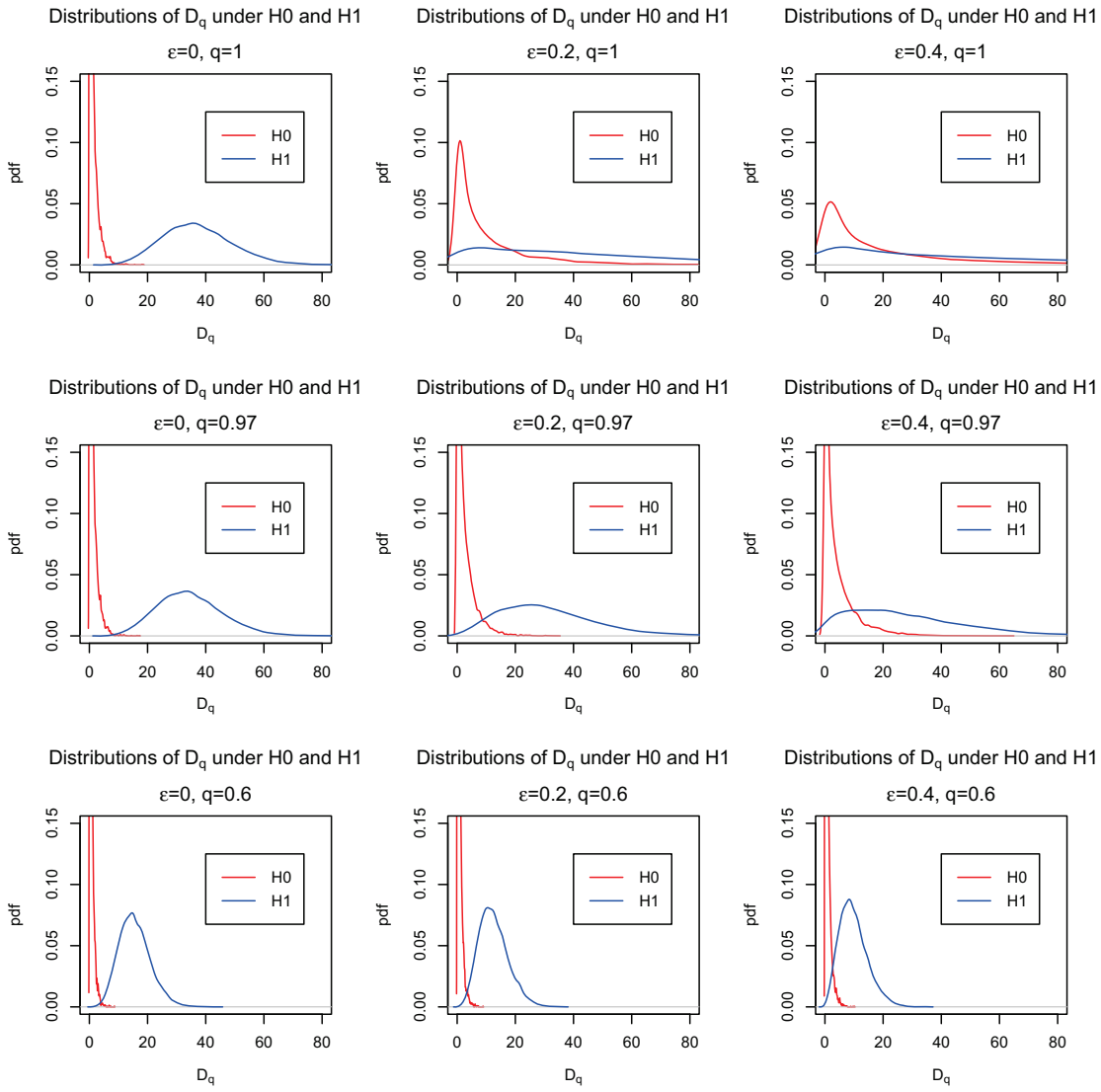


Figure 3.4: Comparison of pdfs of $D_q(\mathbf{x})$ under the null and alternative hypotheses.

distribution or its critical values for different sizes. In order to solve this problem, we need to estimate the critical value from the sample. We propose a bootstrap method for estimating the critical value. It is described as follows. (Suppose we are testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ where θ is the location parameter.)

Step 1: Given a sample $\mathbf{x} = (x_1, ..., x_n)$, we estimate the mean using a robust procedure, e.g., MLq estimate of the sample mean, $\hat{\theta}_q$.

Step 2: Subtract the sample by its estimated mean $\hat{\theta}_q$ and get $\mathbf{x}' = (x_1 - \hat{\theta}_q, ..., x_n - \hat{\theta}_q)$.

Step 3: Perform a bootstrap using $\mathbf{x}'$ and get bootstrap samples $\mathbf{x}'_\mathbf{b}$ for $b = 1, ..., B$.

Step 4: Calculate $D_q(\mathbf{x}'_\mathbf{b})$ for each bootstrap sample and denote each as D_q^b.

Step 5: Calculate the $1 - \alpha$ quantile of D_q^b. Denote it as $\widehat{CV}_\alpha$.

$\widehat{CV}_\alpha$ is our final estimate for the critical value. The rationale behind our bootstrap method is that since we are interested in the null distribution under $H_0 : \theta = 0$, we need to demean the observed sample $\mathbf{x}$ to get a zero mean sample $\mathbf{x}'$. With this zero mean sample $\mathbf{x}'$, we can use the bootstrap to mimic the null distribution. However, since there are usually outliers in the sample, we need to use a robust estimation for the mean. In our case, we adopt the MLqE of the sample. This robust mean helps us to mimic the null distribution.

3.4 Numerical Results and Validation

3.4.1 Simulation

Let us assume f is a normal distribution with a unknown mean θ and a unknown variance σ_f^2. We want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. We simulate data with the sample size $n = 50$ from $h(x; \theta, \epsilon) = (1-\epsilon)\varphi(x; \theta, 1) + \epsilon\varphi(x; \theta, 50)$ where $\varphi(x; \theta, 1)$ corresponds to f. We apply the LqLR (with $q = 1, 0.9, 0.6$), the Wilcoxon test and the sign test on the data. Note that $q = 1$ is essentially the LR (or equivalently, the t test). At different levels of ϵ, we use $h(x; \theta = 0, \epsilon)$ to generate the data 3000 times and calculate the size and then use $h(x; \theta = 0.34, \epsilon)$ to generate the data and calculate the power. The results are shown in Figure 3.5.

In Figure 3.5, let us first note that the size of all tests are successfully controlled at 0.05. At $\epsilon = 0$, the LqLR with $q = 1$ (LR) has the highest power; as we decrease q to 0.9 and 0.6, the power decreases. The Wilcoxon test also has a high power. The sign test has the lowest power. As contamination becomes more serious, i.e., ϵ increases away 0, the LqLR with $q = 1$ (LR) degrades much faster than any other tests. With smaller q's, the LqLRs ($q = 0.9$ and $q = 0.6$) degrade at much slower rates. The Wilcoxon test also degrades slowly. Among all tests, the Lq ratio test with $q = 0.6$ and the sign test have the slowest degradation rates (i.e., flattest curves). By adjusting the tuning parameter q to 0.9, we can beat the Wilcoxon test at mild contamination ($\epsilon < 0.05$). If we change q to 0.6, we can beat the Wilcoxon test

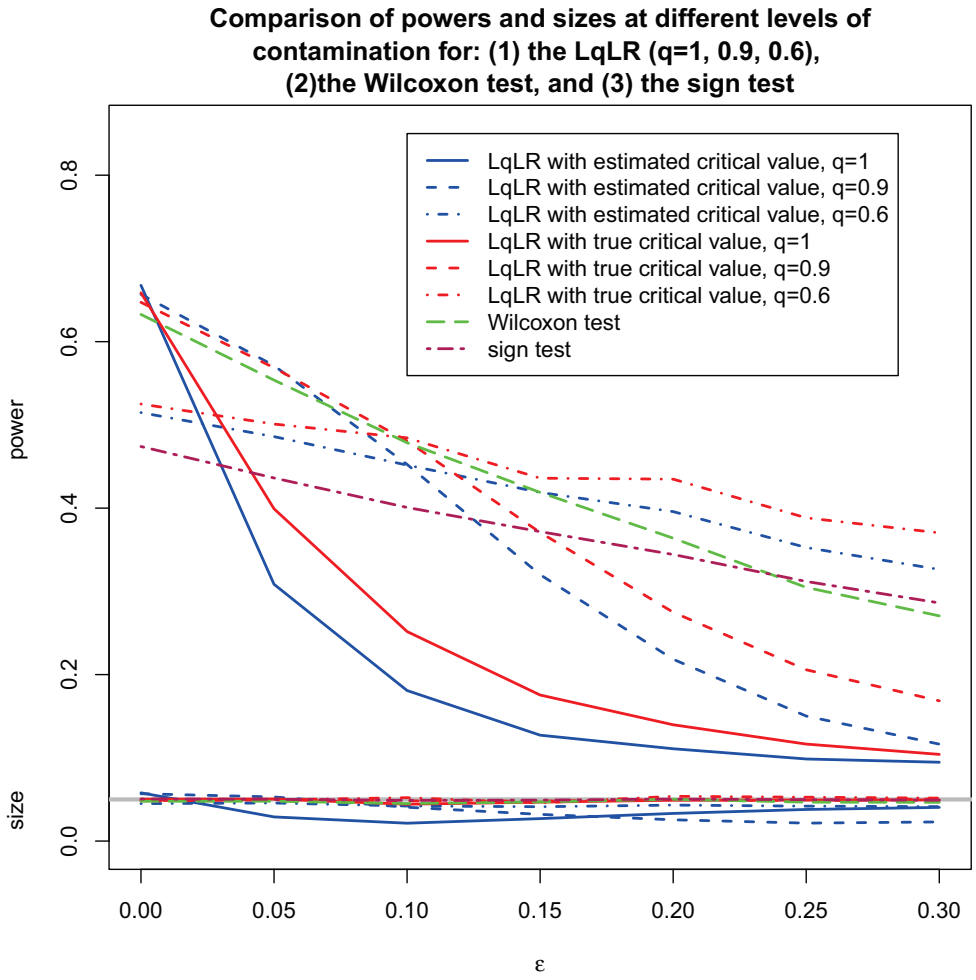


Figure 3.5: Comparison of powers and sizes for the LqLR for $q = 1$ (i.e., the LR or the t test), $q = 0.9$ and $q = 0.6$, the Wilcoxon test and the sign test at different levels of contamination. The blue curves represent the LqLR with estimated critical values. Since we know the true data generating process h, we can simulate the data under h to get the true critical values. We denote the LqLR using the true critical values with red curves. Note that, in practice, it is impossible to know the true data generating process. We present such a scenario only as a benchmark for our proposed method.

at heavy contamination ($\epsilon > 0.15$). Meanwhile, the LqLR with $q = 0.6$ uniformly dominates the sign test at all levels of contamination. Last but not least, the figure also shows that our estimated critical values work well. We only slightly overestimate the critical values, therefore, the powers obtained from the estimated critical values are slightly below the powers obtained from the true critical values.

We see remarkable robustness can be obtained by using the LqLR. The figure also implies that, with an appropriately selected q, it is possible that the LqLR can uniformly beat the Wilcoxon test and the sign test (See Section 3.5 for details). This conjecture is reasonable (and turns out to be true) because, by setting q between 0 and 1, we essentially put more weight on a smaller portion of the data. So the amount of information used in the test becomes smaller and smaller. The extreme case is the sign test which uses only the information of whether each data point is above or below 0. In Section 3.5, we show how the LqLR beats these nonparametric tests.

3.4.2 Real Data

We use a real data example to demonstrate the effectiveness of our proposed method. The data was first presented in [13] and later used in [14]. [13] conducted the experiment to illustrate the effects of optimal isomers of hyoscyamine hydrobromide in producing sleep. There were 10 patients in total. Each patient was given two types of drugs in a randomized order and was asked to record their average sleeping hours

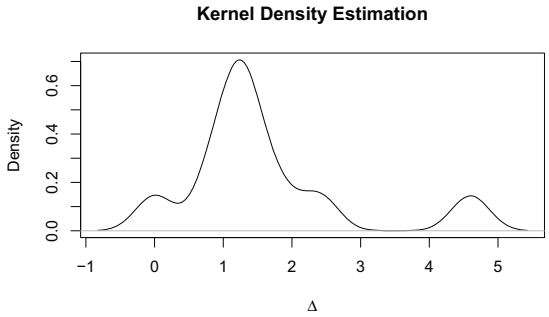


Figure 3.6: Kernel density estimation of the difference in sleep hours gained for the two drugs.

gained for the two drugs. Furthermore, the differences in sleeping hours gained for the two drugs, Δ, are calculated: 1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4. We want to test the null hypothesis that two drugs have the same effect, i.e., $H_0 : u_\Delta = 0$, $H_1 : u_\Delta > 0$, where $u_\Delta = E[\Delta]$.

The importance of this data set is that many statisticians have examined it assuming the normal distribution (including William S. Gosset with his Student's t test). However, the value $\Delta_9 = 4.6$ raises some questions against the normality assumption. A kernel density estimation of Δ (with band width of 0.2755) is presented in Figure 3.6, where we see $\Delta_9 = 4.6$ clearly brings doubt on the normality assumption. For a level of 0.05 test, we can reject H_0 using the t test (equivalent to the LR). If we were to replace the value of 4.6 with 16, then the t test would no longer reject H_0 at the level of 0.05. One may argue that $\Delta_9 = 16$ is an obvious outlier; however, it is counterintuitive that more extreme evidence is favorable to the null hypothesis — no difference in two drugs.

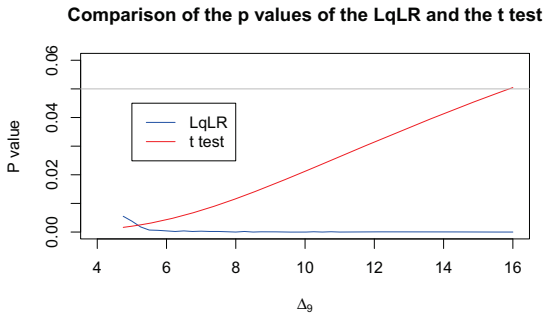


Figure 3.7: Comparison of p values of the LqLR and the t test as functions of Δ_9.

Meanwhile, we apply the LqLR on the data set with $q = 0.85$. In Figure 3.7, we plot the p-value as a function of Δ_9 (which goes from 4.6 to 16) for both the LqLR and the t test. As we can see from the figure, the p-value of the t test gradually increases above 5% as Δ_9 increases. On the other hand, the p-value of the LqLR is well controlled and decreases to 0 as Δ_9 increases. Therefore, the LqLR successfully rejects the null hypothesis with the p-value being consistent with the evidence. Even though the t test (LR) is a special case of the LqLR, by setting $q < 1$, we preserve the efficiency and attain remarkable robustness.

3.5 Selection of q

So far in the chapter, we assume q to be known. However, in practice, we need to pick q for our analysis. In this section, we propose a method for adaptively selecting the tuning parameter q. As we know, the more contamination present, the more protection we need for the power, therefore, the smaller q we should pick. The

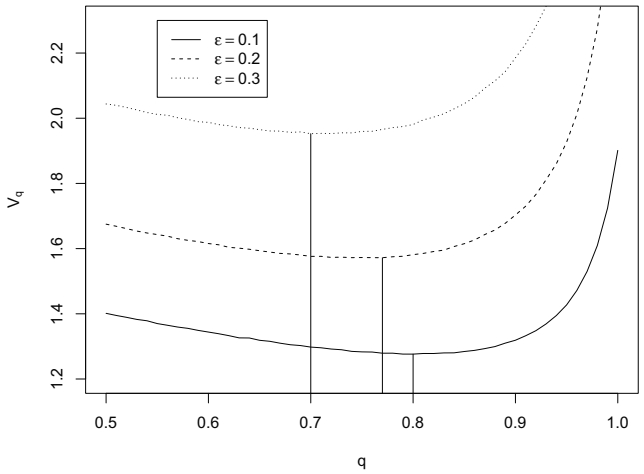


Figure 3.8: $V_q(\theta_0)$ as a function of q at different levels of contamination ratio ϵ.

optimal q we propose is defined as $q_{\text{opt}} = \arg\max_q \Pi$, where Π is the limiting power of the test, i.e., asymptotic power. When testing for the location parameter in the symmetric distribution, we have $\Pi = \Phi(\frac{\delta}{V_q(\theta_0)} - u_\alpha)$. Since this is a monotonic function in $V_q(\theta_0)$, our optimal q is given by $q_{\text{opt}} = \arg\min_q V_q(\theta_0)$.

In Figure 3.8, we plot the relationship between $V_q(\theta_0)$ and q at different levels of contamination using the same set up in the previous section. We can clearly see that the optimal q is between 0.6 to 0.9 for these contamination levels. As expected, the higher the contamination ratio is, the lower the optimal q is.

In practice, we do not have $V_q(\theta_0)$. We can replace it with the empirical version of this quantity. The data-adaptive estimation for the tuning parameter is given by

$$\hat{q} = \arg\min_q \frac{\frac{1}{n}\sum_{i=1}^n \psi_q(x_i;\hat{\theta}_q)^2}{[\frac{1}{n}\sum_{i=1}^n \psi'_q(x_i;\hat{\theta}_q)]^2}.$$

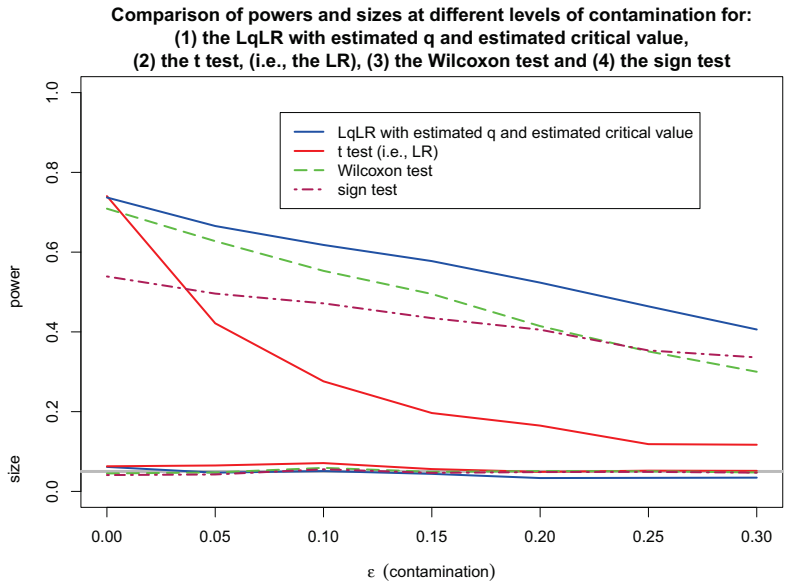


Figure 3.9: Comparison of the powers and sizes of: 1). the LqLR with the estimated q and the estimated critical value; 2) the t test, i.e., the LR; 3) the Wilcoxon test and 4) the sign test at different levels of contamination.

We now provide a simulation study of the LqLR using estimated q and estimated critical value. We adopt the same set up from the previous section (Section 3.4.1). Using 2000 Monte Carlo iterations, we compare the power and size of the LqLR, the LR, the Wilcoxon test and the sign test at different levels of contamination. The results are demonstrated in Figure 3.9. We can clearly see the advantage of the LqLR (with estimated q and estimated critical value) over other tests. Not only does the LqLR degrade very slowly, it also holds the highest power among all other tests. Note that the sizes have been successfully controlled at 5%. In Figure 3.9, at zero contamination (i.e., $\epsilon = 0$), the LR has the highest power. The LqLR has almost the same power (only slightly less than the LR). The Wilcoxon and the sign tests have

the third and the fourth highest powers, but not comparable to the two likelihood ratio tests. As the contamination becomes more serious (i.e., ϵ increases away 0), the log-likelihood degrades the fastest. Its power quickly drops below all other tests. The Wilcoxon test and the sign test both show good robustness and their powers degrade at much slower rates. However, the LqLR shows a remarkable robustness. It degrades slower than the Wilcoxon test (i.e., the blue curve is flatter than the green curve) and only slightly faster than the sign test (i.e., the blue curve is steeper than the maroon curve). Since the power of the LqLR at $\epsilon = 0$ is above that of the Wilcoxon test and the sign test, the power of the LqLR dominates both the Wilcoxon test and the sign test at all levels of contamination. This implies that, not only can Lq-likelihood preserve efficiency almost perfectly at $\epsilon = 0$, it also obtains robustness comparable to these nonparametric tests which are known to be very robust. We conclude that, by losing a little efficiency at $\epsilon = 0$, we have traded for great robustness at $\epsilon > 0$. Our LqLR can be considered as a combination of the LR (at $\epsilon = 0$) and the nonparametric tests (at $\epsilon > 0$). The reason our test beats nonparametric tests uniformly is that we can control the amount of information to use by selecting q, whereas the Wilcoxon test always uses the rank information, and the sign test always uses the information about whether each data point is below or above the hypothesized mean.

Meanwhile, we also plot the histograms of the estimated q at different levels of contamination in Figure 3.10. We see that as we get more serious contamination, the estimated q tends to be smaller. In our experiment, we limit the smallest q to be 0.5,

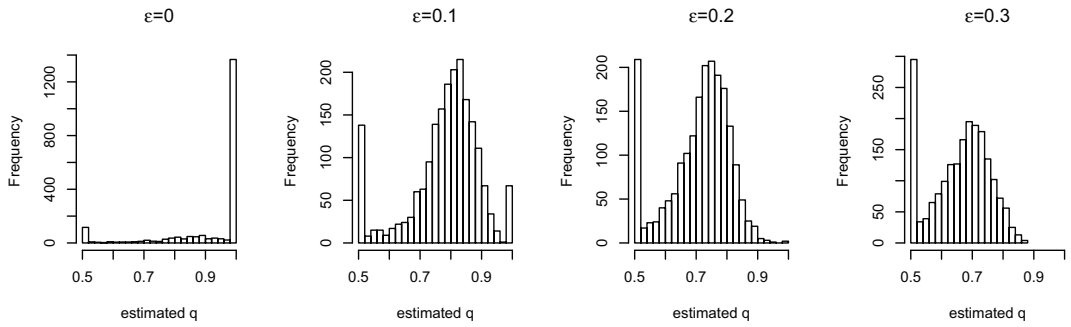


Figure 3.10: Histogram of the estimated q at different levels of contamination.

which is very similar to the case of testing based on minimum Hellinger distance [15]. Whenever our estimated q drops below 0.5, we use 0.5 instead. The reason for this censoring is that we have not understood the case of $q < 0.5$ very well, which is an interesting topic for future research.

3.6 Conclusion

In this chapter, we have proposed a robust testing procedure — the Lq-likelihood ratio test (LqLR) — and demonstrated its advantage over the LR, the Wilcoxon test, and the sign test under the gross error model for testing the location parameter of a symmetric distribution. We prove the LqLR's robustness advantages by deriving the asymptotic distribution. We further accompany our analytical study with numerical comparisons.

Our LqLR can be considered as a bridge connecting the LR and the nonparametric tests such as the Wilcoxon test and the sign test. By changing the tuning parameter

q, we can control the information used in the hypothesis testing. The LR uses the full information of all data points and gives all data points equal weights. The Wilcoxon test takes only the rank information, and therefore becomes extremely robust at the cost of wasting much information. Our LqLR gives each data point a weight as a function of its likelihood and q. Therefore, the data points consistent with the "idealized" model are given higher weights whereas data points inconsistent with the "idealized" model are partially ignored.

To the extent that the robustness of the Wilcoxon test (minimum asymptotic relative efficiency (ARE) of the Wilcoxon test vs the t test is 0.864) suggests that the Wilcoxon test should be the default test of choice (rather than "use Wilcoxon if there is evidence of non-normality," the default position should be "use Wilcoxon unless there is good reason to believe the normality assumption"), these new results in this chapter suggest that the LqLR test should become the new default go-to test for practitioners everywhere!

Even though our test shows remarkable robustness over other tests, there are still many directions for future research. For example, the investigation of the LqLR's properties under the asymmetric distribution is an important topic. Meanwhile, better estimation procedures for the critical value and q are needed. We have shown that our estimate of the critical value performs decently, but there is clearly a gap between the power obtained from the true critical values and the power obtained from the estimated critical values. Filling in that gap is a challenging task for the future.

The estimation of q also leaves many directions for future research. We could develop a much more robust procedure for selecting q. Finally, all the conclusions in this chapter are for the location parameter; we suspect the same effect will hold for the scale parameter, which is also an important direction for future research. However, the contrast function $\psi_q(x;\theta)$ for the scale parameter is significantly different from that of the location parameter. Therefore, handling the scale parameter is much more challenging.

Chapter 4

MLqE for Mixture Models

In this chapter, we introduce a maximum Lq-likelihood estimation (MLqE) of mixture models using our proposed expectation maximization (EM) algorithm, namely the EM algorithm with Lq-likelihood (EM-Lq). Properties of the MLqE obtained from the proposed EM-Lq are studied through simulated mixture model data. Compared with the maximum likelihood estimation (MLE) which is obtained from the EM algorithm, the MLqE provides a more robust estimation against outliers for small sample sizes. In particular, we study the performance of the MLqE in the context of the gross error model, where the true model of interest is a mixture of two normal distributions, and the contamination component is a third normal distribution with a large variance. A numerical comparison between the MLqE and the MLE for this gross error model is presented in terms of Kullback Leibler (KL) distance and relative efficiency.

4.1 MLqE of Mixture Models

We now look at the problem of estimating mixture models. A mixture model is defined as $f(x) = \sum_{j=1}^{k} \pi_j f_j(x;\theta_j)$. Unlike the exponential family which is proved to be closed under the escort transformation (equation (2.4)), the mixture model family is not closed under such a transformation. For example, consider a mixture model with the complexity $k = 2$. The escort transformation with $1/q = 2$ of this distribution is

$$\begin{aligned} f(x)^{(1/q)} &\propto (\pi_1\varphi_1(x) + \pi_2\varphi_2(x))^2 \\ &= \pi_1^2\varphi_1(x)^2 + \pi_2^2\varphi_2(x)^2 + 2\pi_1\pi_2\varphi_1(x)\varphi_2(x), \end{aligned}$$

which is a mixture model with three components.

More generally, suppose $f_0 \in \mathcal{F}$, where $\mathcal{F}$ is a mixture model family with complexity k. Since $f_0^{(1/q)} \notin \mathcal{F}$, we know that

$$f_0^{(1/q)} \neq \tilde{g} := \arg\max_{g\in\mathcal{F}} E_{f_0} L_q(g(X)),$$

where $\tilde{g}$ can be considered as the projection of $f_0^{(1/q)}$ onto $\mathcal{F}$. Again, the MLqE of mixture models brings more bias to the estimate. This time, the new bias is a model bias as opposed to the estimation bias which we have discussed in Chapter 2. When estimating mixture models using MLqE, we encounter two types of bias: estimation

bias and model bias. The distortion parameter q now adjusts both of them. This idea is illustrated in Figure 4.1.

There is a simple way to partially correct the bias. Since we know that the MLqE is unbiased for the escort distribution of the true distribution, after we obtain the MLqE from data, $\hat{f}_{\mathrm{ML}q\mathrm{E}}$, we can apply to it a power transformation $g = \hat{f}^q_{\mathrm{ML}q\mathrm{E}} / \int \hat{f}^q_{\mathrm{ML}q\mathrm{E}} dx$ to get a less biased estimate. However, this only partially corrects the bias since the projection from the escort distribution onto the mixture model family cannot be recovered by this transformation.

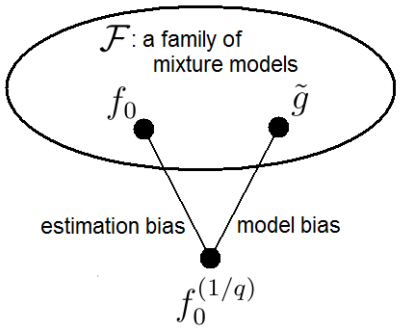


Figure 4.1: Illustration of the MLqE of mixture models: the MLqE of mixture models with correctly specified models in the usual case.

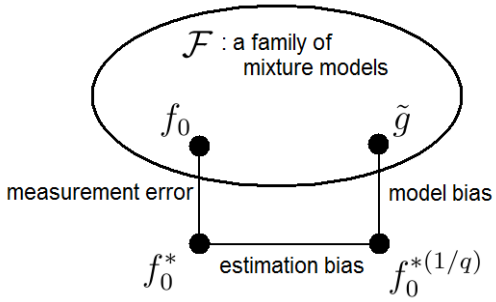


Figure 4.2: Illustration of the MLqE of mixture models: the MLqE of non-measurement error components f_0 within the gross error model f_0^* using the misspecified model.

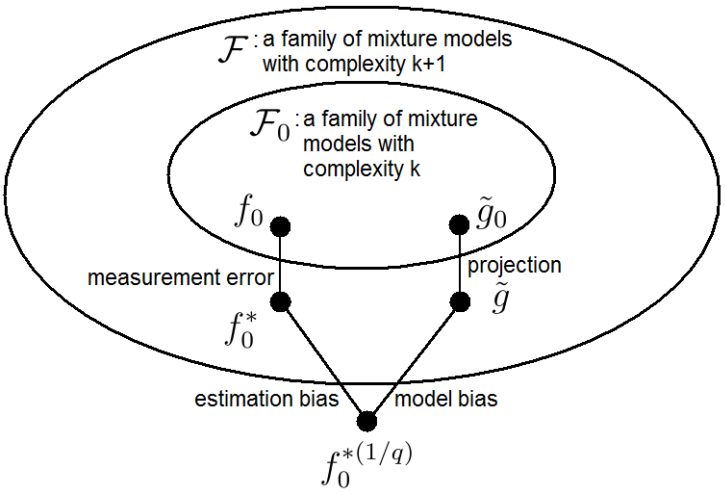


Figure 4.3: Illustration of the MLqE of mixture models: the MLqE of non-measurement error components f_0 within the gross error model f_0^* using the correctly specified model.

Because the MLqE has the desirable property of being robust against outliers, we introduce the gross error model to evaluate the MLqE's performance. A gross error model is defined as $f_0^*(x) = (1-\epsilon)f_0(x) + \epsilon f_{\text{err}}(x)$, where f_0 is a mixture model with complexity k, f_{err} can be considered as a measurement error component, and ϵ is the contamination ratio. Hence, f_0^* is also a mixture model with complexity $k+1$. The gross error density f_0^* can be considered as a small deviation from the target density f_0. In order to build an estimator for f_0 that is robust against f_{err}, we apply the MLqE. Generally, there are two ways to apply the MLqE in this situation.

First, we can directly use a mixture model with complexity k to estimate f_0 based on data from f_0^*. We call this approach the direct approach. This time the model is more complex than before. The idea is illustrated in Figure 4.2. Suppose $\mathcal{F}$ is a mixture model family with complexity k, and $f_0 \in \mathcal{F}$, $f_0^* \notin \mathcal{F}$, $f_0^{*(1/q)} \notin \mathcal{F}$. We obtain

the MLqE of $f_0(x)$, $\tilde{g}$, by

$$f_0^{*(1/q)} \neq \tilde{g} := \arg\max_{g \in \mathcal{F}} E_{f_0^*} L_q(g(X)).$$

Here we use the estimation bias and the model bias to offset the measurement error effect on f_0. Please note that this approach is essentially an estimation under the misspecified model.

The second approach is that we use a mixture model with complexity $k+1$ to estimate f_0^* and project the estimate to the k component mixture model family by removing the largest variance component (i.e., the measurement error component) and normalizing the weights. We call this approach the indirect approach. The projected model is our estimate for f_0. In this case, we essentially treat the parameters of the measurement error component as nuisance parameters. This idea is illustrated in Figure 4.3. In Figure 4.3, $\tilde{g}$ is our estimate of f_0^*. And $\tilde{g}_0$, the projection of $\tilde{g}$ onto $\mathcal{F}_0$, is our estimate of f_0. This approach is an estimation conducted under the correctly specified model. Although the model is correctly specified, we may have higher estimation variance as we estimate more parameters.

In this chapter, we will study the MLqE using the above two approaches.

Please note that, when $q \neq 1$, the MLqE is an inconsistent estimator. [1] let $q \to 1$ as $n \to \infty$ in order to force the consistency. In our case, we allow the MLqE to be inconsistent because our data is contaminated. We are no longer after the

true underlying distribution f_0^* that generates the data, but are more interested in estimating the non-measurement error components f_0 using the contaminated data. Since the goal is not to estimate f_0^*, being consistent will not help the estimator in terms of robustness.

4.2 EM Algorithm with Lq-Likelihood

We now propose a variation of the EM algorithm — the expectation maximization algorithm with Lq-likelihood (EM-Lq), which gives the local maximum Lq-likelihood. Before introducing our EM-Lq, let us briefly review the rationale of the EM. Throughout this chapter, we use X, Z, $\mathbf{Z}$ for random variables and vectors, and x, z, $\mathbf{z}$ for realizations.

4.2.1 Why Does the EM Algorithm Work

The EM algorithm is an iterative method for finding a local maximum likelihood by making use of observed data X and missing data Z. The rationale behind the EM is that

$$\sum_{i=1}^{n} \log p(x_i; \Psi) = \underbrace{\sum_{i=1}^{n} E_{\Psi^{\text{old}}}[\log p(X, Z; \Psi)|X = x_i]}_{J(\Psi, \Psi^{\text{old}})} \underbrace{- \sum_{i=1}^{n} E_{\Psi^{\text{old}}}[\log p(Z|X; \Psi)|X = x_i]}_{K(\Psi, \Psi^{\text{old}})},$$

where $J(\Psi, \Psi^{\text{old}})$ is the expected complete log likelihood, and $K(\Psi, \Psi^{\text{old}})$ takes its minimum at $\Psi = \Psi^{\text{old}}$ and $\frac{\partial}{\partial \Psi} K(\Psi, \Psi^{\text{old}})\big|_{\Psi=\Psi^{\text{old}}} = 0$. Standing at the current estimate Ψ^{old}, to climb uphill on $\sum_{i=1}^{n} \log p(x_i; \Psi)$ only requires us to climb J, and K will automatically increase. Meanwhile, the incomplete log likelihood and the expected complete log likelihood share the same derivative at $\Psi = \Psi^{\text{old}}$, i.e.,

$$\frac{\partial}{\partial \Psi} \sum_{i=1}^{n} \log p(x_i; \Psi)\Big|_{\Psi=\Psi^{\text{old}}} = \frac{\partial}{\partial \Psi} J(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}}. \tag{4.1}$$

This is also known as the minorization-maximization algorithm (MM). A detailed explanation of the algorithm can be found in [16]. Our algorithm presented in the next section is essentially built on [16] with variation made for the Lq-likelihood.

4.2.2 EM Algorithm with Lq-Likelihood

Having the idea of the traditional EM in mind, let us maximize the Lq-likelihood $\sum_{i=1}^{n} L_q(p(x_i; \Psi))$ in a similar fashion. For any two random variables X and Z, we have

$$L_q(p(X; \Psi)) = L_q\Big(\frac{p(X, Z; \Psi)}{p(Z|X; \Psi)}\Big) = \frac{L_q(p(X, Z; \Psi)) - L_q(p(Z|X; \Psi))}{p(Z|X; \Psi)^{1-q}},$$

where we have used $L_q(a/b) = [L_q(a) - L_q(b)]/b^{1-q}$ (Lemma 7.6.1, part (iii) in Chapter 7). Applying the above equation on data $x_1, ..., x_n$, and taking expectation (under

Ψ^{old}) given observed data $x_1, ..., x_n$, we have

$$\sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[L_q(p(X;\Psi))\Big|X = x_i\Big] =$$
$$\sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(p(X,Z;\Psi)) - L_q(p(Z|X;\Psi))}{p(Z|X;\Psi)^{1-q}}\Big|X = x_i\Big],$$
$$\sum_{i=1}^{n} L_q(p(x_i;\Psi)) =$$
$$\sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\Big(\frac{p(Z|X;\Psi^{\text{old}})}{p(Z|X;\Psi)}\Big)^{1-q}\Big(\frac{L_q(p(X,Z;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}} - \frac{L_q(p(Z|X;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}}\Big)\Big|X = x_i\Big],$$

where we multiply and divide $P(Z|X,\Psi^{\text{old}})^{1-q}$ in the numerator and the denominator.

Define

$$A(\Psi,\Psi^{\text{old}}) = \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(p(X,Z;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}} - \frac{L_q(p(Z|X;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}}\Big|X = x_i\Big],$$
$$B(\Psi,\Psi^{\text{old}}) = \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(p(X,Z;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}}\Big|X = x_i\Big],$$
$$C(\Psi,\Psi^{\text{old}}) = -\sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(p(Z|X;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}}\Big|X = x_i\Big],$$
$$\Rightarrow A(\Psi,\Psi^{\text{old}}) = B(\Psi,\Psi^{\text{old}}) + C(\Psi,\Psi^{\text{old}}). \tag{4.2}$$

Based on the definitions above, we have the following theorems.

Theorem 4.2.1. $C(\Psi,\Psi^{\text{old}})$ takes its minimum at $\Psi = \Psi^{\text{old}}$, i.e., $C(\Psi^{\text{old}},\Psi^{\text{old}}) = \min_{\Psi} C(\Psi,\Psi^{\text{old}})$.

Proof.

$$\begin{aligned} C(\Psi^{\text{old}}, \Psi^{\text{old}}) - C(\Psi, \Psi^{\text{old}}) &= \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[L_q\Big(\frac{p(Z|X;\Psi)}{p(Z|X;\Psi^{\text{old}})}\Big)\Big|X = x_i\Big] \\ &\leq \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{p(Z|X;\Psi)}{p(Z|X;\Psi^{\text{old}})} - 1\Big|X = x_i\Big] \\ &= \sum_{i=1}^{n}\sum_{z}\Big(\frac{p(z|x_i;\Psi)}{p(z|x_i;\Psi^{\text{old}})} - 1\Big)p(z|x_i;\Psi^{\text{old}}) = 0, \end{aligned}$$

where the inequality comes from the fact that $L_q(u) \leq u - 1$ (Lemma 7.6.1, part (iv) in Chapter 7). The above inequality becomes equality only when $\Psi = \Psi^{\text{old}}$. □

Theorem 4.2.2. When A, B and C are differentiable with respect to Ψ, we have

$$\begin{aligned} &\frac{\partial}{\partial\Psi}C(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} = 0, \\ &\frac{\partial}{\partial\Psi}A(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} = \frac{\partial}{\partial\Psi}B(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}}. \end{aligned} \tag{4.3}$$

Proof. The first part is a direct result from Theorem 4.2.1. By equation (4.2) and the first part of the theorem, we have the second part. □

Comparing equation (4.3) with equation (4.1), we can think of B as a proxy of the complete Lq-likelihood (i.e., J), A as a proxy of the incomplete Lq-likelihood, and C as a proxy of K.

We know that A is only an approximation of $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$ due to the factor

of $(p(Z|X;\Psi^{\text{old}})/p(Z|X;\Psi))^{1-q}$. However, at $\Psi = \Psi^{\text{old}}$, we do have

$$A(\Psi,\Psi^{\text{old}})\big|_{\Psi=\Psi^{\text{old}}} = \sum_{i=1}^{n} L_q(p(x_i;\Psi))\big|_{\Psi=\Psi^{\text{old}}}. \tag{4.4}$$

A will be a good approximation of $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$ because: (1) within a small neighborhood $N_r(\Psi^{\text{old}}) = \{\Psi : d(\Psi,\Psi^{\text{old}}) < r\}$, $p(Z|X;\Psi^{\text{old}})/p(Z|X;\Psi)$ is approximately 1; (2) due to the transformation $y = x^{1-q}$, $(p(Z|X;\Psi^{\text{old}})/p(Z|X;\Psi))^{1-q}$ gets pushed toward 1 even further when q is close to 1; and (3) even if $(p(Z|X;\Psi^{\text{old}})$ $/p(Z|X;\Psi))^{1-q}$ is far from 1, because we sum over all the x_i's, we still average out these poorly approximated data points.

Given that C achieves a minimum at Ψ^{old}, starting at Ψ^{old} and maximizing A requires only maximizing B. In order to take advantage of this property, we use A to approximate $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$ at each iteration, and then maximize B to maximize A, and eventually to maximize $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$. B is usually easy to maximize. Based on this idea, we build our EM-Lq as follows:

- 1. E step: Given Ψ^{old}, calculate B.
- 2. M step: Maximize B and obtain $\Psi^{\text{new}} = \arg\max_{\Psi} B(\Psi,\Psi^{\text{old}})$.
- 3. If Ψ^{new} converges, we terminate the algorithm. Otherwise, we set $\Psi^{\text{old}} = \Psi^{\text{new}}$, and return to step 1.

4.2.3 Monotonicity and Convergence

In this section, we will discuss the monotonicity and the convergence of the EM-Lq algorithm. We start with the following theorem.

Theorem 4.2.3. For any Ψ, we have the lower bound of the Lq-likelihood function

$$\sum_{i=1}^{n} L_q(p(x_i;\Psi)) \geq B(\Psi,\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}). \tag{4.5}$$

When $\Psi = \Psi^{\text{old}}$, we have

$$\sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{old}})) = B(\Psi^{\text{old}},\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}).$$

Proof. See Chapter 7 for proof. □

From Theorem 4.2.3, we know that, at each M step, as long as we can find Ψ^{new} that increases B, i.e., $B(\Psi^{\text{new}},\Psi^{\text{old}}) > B(\Psi^{\text{old}},\Psi^{\text{old}})$, we can guarantee that the Lq-likelihood will also increase, i.e., $\sum_{i=1}^{n} L_q(x_i;\Psi^{\text{new}}) > \sum_{i=1}^{n} L_q(x_i;\Psi^{\text{old}})$. This is because

$$\begin{aligned}
\sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{new}})) &\geq B(\Psi^{\text{new}},\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}) \\
&> B(\Psi^{\text{old}},\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}) \\
&= \sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{old}})).
\end{aligned}$$

Thus, we have proved the monotonicity of our EM-Lq algorithm.

Based on Theorem 4.2.3, we can further derive the following theorem.

Theorem 4.2.4. For our EM-Lq algorithm, when A, B and the Lq-likelihood are differentiable with respect to Ψ, it holds that

$$\frac{\partial}{\partial\Psi}\sum_{i=1}^{n} L_q(p(x_i;\Psi))\Big|_{\Psi=\Psi^{\text{old}}} = \frac{\partial}{\partial\Psi}B(\Psi,\Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} = \frac{\partial}{\partial\Psi}A(\Psi,\Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}},$$
$$\sum_{i=1}^{n} L_q(P(x_i;\Psi))\Big|_{\Psi=\Psi^{\text{old}}} = A(\Psi,\Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}}.$$

Proof. See Chapter 7 for proof. □

It becomes clear that A is not only just a good approximation of, but also the first order approximation of, $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$.

One good thing following from the property of the first order approximation is that, when we have a fixed point, meaning that $A(\Psi^{\text{old}},\Psi^{\text{old}}) = \max_\Psi A(\Psi,\Psi^{\text{old}})$, then we know $\frac{\partial}{\partial\Psi}A(\Psi,\Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} = \frac{\partial}{\partial\Psi}\sum_{i=1}^{n} L_q(p(x_i;\Psi))\Big|_{\Psi=\Psi^{\text{old}}} = 0$, which means that $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$ takes its local maximum at the same place that $A(\Psi,\Psi)$ does. So as long as we achieve the maximum of A, we simultaneously maximize the incomplete Lq-likelihood $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$.

By Theorem 4.2.4, we know that, as long as $\frac{\partial}{\partial\Psi}B(\Psi,\Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} \neq 0$, we can always find a Ψ^{new}, such that $\sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{new}})) > \sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{old}}))$. Hence our EM-L$q$ can be considered as a generalized EM algorithm (GEM) for Lq-likelihood. [17] has proved the convergence of the GEM from a pure optimization approach (Global

Convergence Theorem, Theorem 1 and Theorem 2 of [17] pp. 97 - 98), which we can directly use to prove the convergence of the EM-Lq.

In our simulation results, the converging point of the EM-Lq is always the same as the true maximizer of the Lq-likelihood which is obtained from the optimization package fmincon() in Matlab. We also try to move a small step away from the solution given by the EM-Lq to check whether the Lq-likelihood decreases. This shows that a small step in any direction will cause the Lq-likelihood to decrease, which numerically demonstrates that the solution is a local maximizer.

4.3 EM-Lq Algorithm for Mixture Models

4.3.1 EM-Lq for Mixture Models

Returning to our mixture model, suppose the observed data $x_1, ..., x_n$ are generated from a mixture model $f(x; \Psi) = \sum_{j=1}^{k} \pi_j f_j(x; \theta_j)$ with parameter $\Psi = (\pi_1, \ldots, \pi_{k-1}, \theta_1, \ldots, \theta_k)$. The missing data are the component labels $[\mathbf{z_1}, ..., \mathbf{z_n}]$, where $\mathbf{z_i} = (z_{i1}, ..., z_{ik})$ is a k dimensional component label vector with each element z_{ij} being 0 or 1 and $\sum_{j=1}^{k} z_{ij} = 1$.

In this situation, we have

$$p(x, \mathbf{z}; \Psi) = \prod_{j=1}^{k} (\pi_j f_j(x; \theta_j))^{z_j}, \tag{4.6}$$

$$p(\mathbf{z}|x; \Psi) = \prod_{j=1}^{k} p(z_j|x; \Psi)^{z_j} = \prod_{j=1}^{k} \Big(\frac{\pi_j f_j(x; \theta_j)}{f(x; \Psi)}\Big)^{z_j}, \tag{4.7}$$

where x is an observed data point, and $\mathbf{z} = (z_1, ..., z_k)$ is a component label vector. Substituting these into B and reorganizing the formula, we have

Theorem 4.3.1. In the mixture model case, B can be expressed as

$$B(\Psi, \Psi^{\text{old}}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_j(x_i, \Psi^{\text{old}})^q L_q(\pi_j f_j(x_i; \theta_j)),$$

where $\tau_j(x_i, \Psi^{\text{old}}) = E_{\Psi^{\text{old}}}[Z_{ij}|X = x_i]$, i.e., the soft label in the traditional EM.

Proof. See Chapter 7 for proof. □

We define new binary random variables $\tilde{Z}_{ij}$ whose expectation is $\tilde{\tau}_j(x_i, \Psi^{\text{old}}) = E_{\Psi^{\text{old}}}[\tilde{Z}_{ij}|X = x_i] = E_{\Psi^{\text{old}}}[Z_{ij}|X = x_i]^q$. $\tilde{Z}_{ij}$ can be considered as a distorted label as its probability distribution is distorted (i.e., $P_{\Psi^{\text{old}}}(\tilde{Z}_{ij} = 1|x_i) = P_{\Psi^{\text{old}}}(Z_{ij} = 1|x_i)^q$). Please note that, for $\tilde{Z}_{ij}$, we no longer have $\sum_{j=1}^{k} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) = 1$. After the replacement, B becomes

$$B(\Psi, \Psi^{\text{old}}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) L_q(\pi_j f_j(x_i; \theta_j)).$$

To maximize B, we apply the first order condition and obtain the following theorem.

Theorem 4.3.2. The first order condition of B with respect to θ_j and π_j yields

$$0 = \frac{\partial}{\partial \theta_j} B(\Psi, \Psi^{\text{old}}) \Rightarrow 0 = \sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \frac{\frac{\partial}{\partial \theta_j} f_j(x_i; \theta_j)}{f_j(x_i; \theta_j)} f_j(x_i; \theta_j)^{1-q}, \tag{4.8}$$

$$0 = \frac{\partial}{\partial \pi_j} B(\Psi, \Psi^{\text{old}}) \Rightarrow \pi_j \propto \Big[\sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) f_j(x_i; \theta_j)^{1-q} \Big]^{\frac{1}{q}}. \tag{4.9}$$

Proof. See Chapter 7 for proof. □

Recall that the M step in the traditional EM solves a similar set of equations,

$$0 = \frac{\partial}{\partial \theta_j} J(\Psi, \Psi^{\text{old}}) \Rightarrow 0 = \sum_{i=1}^{n} \tau_j(x_i, \Psi^{\text{old}}) \frac{\frac{\partial}{\partial \theta_j} f_j(x_i; \theta_j)}{f_j(x_i; \theta_j)}, \tag{4.10}$$

$$0 = \frac{\partial}{\partial \pi_j} J(\Psi, \Psi^{\text{old}}) \Rightarrow \pi_j \propto \sum_{i=1}^{n} \tau_j(x_i, \Psi^{\text{old}}). \tag{4.11}$$

Comparing equations (4.10) and (4.11) with equations (4.8) and (4.9), we see that (1) θ_j^{new} of the EM-Lq satisfies a weighted likelihood equation, where the weights contain both the distorted soft label $\tilde{\tau}_j(x_i, \Psi^{\text{old}})$ and the power transformation of the individual component density function, $f_j(x_i; \theta_j)^{1-q}$; and (2) π_j is proportional to the summation of the distorted soft label $\tilde{\tau}_j(x_i, \Psi^{\text{old}})$ adjusted by the individual density function.

4.3.2 EM-Lq for Gaussian Mixture Models

Consider a Gaussian mixture model with parameter $\Psi = (\pi_1, \ldots, \pi_{k-1}, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2)$. At each E step, we calculate $\tilde{\tau}_j(x_i, \Psi^{\text{old}}) = \left[\frac{\pi_j^{\text{old}} \varphi(x_i; \mu_j^{\text{old}}, \sigma_j^{2\text{old}})}{f(x_i, \Psi_{\text{old}})}\right]^q$. At each M step, we solve equations (4.8) and (4.9) to yield

$$\mu_j^{\text{new}} = \frac{1}{\sum_{i=1}^n \tilde{w}_{ij}} \sum_{i=1}^n \tilde{w}_{ij} x_i, \tag{4.12}$$

$$\sigma_j^{2\text{new}} = \frac{1}{\sum_{i=1}^n \tilde{w}_{ij}} \sum_{i=1}^n \tilde{w}_{ij} (x_i - \mu_j^{\text{new}})^2, \tag{4.13}$$

$$\pi_j^{\text{new}} \propto \Big[\sum_{i=1}^n \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \varphi(x_i; \mu_j^{\text{new}}, \sigma_j^{2\text{new}})^{1-q}\Big]^{\frac{1}{q}},$$

where $\tilde{w}_{ij} = \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \varphi(x_i; \mu_j^{\text{new}}, \sigma_j^{2\text{new}})^{1-q}$. The same iterative re-weighting algorithm designed for solving equations (2.2) and (2.3) can be used to solve equations (4.12) and (4.13). Details of the algorithm are shown in Chapter 7.

At each M step, we can replace $\tilde{w}_{ij}$ with $\tilde{w}_{ij}^* = \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \varphi(x_i; \mu_j^{\text{old}}, \sigma_j^{2\text{old}})^{1-q}$, which only depends on the Ψ^{old}, to improve the efficiency of the algorithm. Thus we can avoid the re-weighting algorithm at each M step. This replacement will simplify the EM-Lq algorithm significantly. We have done simulation demonstrating that this modified version of the algorithm also gives the same solutions as the original EM-Lq algorithm.

4.3.3 Convergence Speed

We present a preliminary comparison of the convergence speeds of the EM-Lq and the EM algorithm using a Gaussian Mixture Model with complexity of 2 (2GMM), $f(x) = 0.4\varphi(x;1,2) + 0.6\varphi(x;5,2)$, whose two components are poorly separated. Surprisingly, in this case, the convergence of the EM-Lq is on average slightly faster than that of the EM.

The comparison of the convergence speed is based on r, which is defined as

$$r = \frac{\|\Psi^{(k)} - \Psi^{(k-1)}\|}{\|\Psi^{(k-1)} - \Psi^{(k-2)}\|},$$

where k is the last iteration of the EM-Lq or the EM algorithm. The smaller r is, the faster the convergence is.

We simulate 1000 data sets according to the 2GMM, use the EM-Lq ($q = 0.8$) and the EM to fit the data, and record the convergence speed difference $r_{\mathrm{ML}q\mathrm{E}} - r_{\mathrm{MLE}}$. The average convergence speed difference is -0.012 with a standard error of 0.002, which means the negative difference in the convergence speed is statistically significant. We note, however, that comparing convergence speed can be misleading in such multi-modal situations.

However, if we change the 2GMM to a gross error model of 3GMM: $f(x) = 0.4(1-\epsilon)\varphi(x;1,2) + 0.6(1-\epsilon)\varphi(x;5,2) + \epsilon\varphi(x;3,40)$, where the third component is an outlier component, and still use a 2GMM to fit, the comparison of the convergence speed

becomes unclear. We have not yet fully understood the properties of the convergence speed for the EM-Lq. However, we do believe the convergence speed is important, and is an interesting topic for future research.

The fact that the convergence of the EM-Lq is a little faster (in this particular case, at least) than that of the EM is closely related to the concept of the information ratio mentioned in [7] and [18], where the convergence speed is connected to the missing information ratio. In Lq-likelihood, since the two in-separable components are pushed apart by the weights $\tilde{w}_{ij}$, the corresponding concept of the missing information ratio for the Lq-likelihood must be relatively lower, thus, we have a faster convergence.

Although the convergence speed is faster for our example for the EM-Lq, it is not necessary that the EM-Lq takes less computer time than the EM. This is because, at each M step in the EM-Lq, we need to do another iterative algorithm to obtain Ψ^{new} (i.e., the algorithm explained in Chapter 7), whereas the EM needs only one step to obtain the new parameter estimate.

The advantage of the convergence speed of the EM-Lq has been hinted at another algorithm called q-Parameterized Deterministic Annealing EM algorithm (q-DAEM) previously proposed by [19] in the signal processing and statistical mechanics context. The q-DAEM can successfully maximize the log-likelihood at a faster convergence speed, by using a different but similar M step as in our EM-Lq. Their M step includes setting $q > 1$ and $\beta > 1$ and dynamically pushing $q \to 1$ and $\beta \to 1$ (β is an additional parameter for their deterministic annealing procedure). On the other hand, our EM-

Lq maximizes the Lq-likelihood with a fixed $q < 1$. Although the objective functions are different for these two algorithms, it is obvious that the advantages in terms of the convergence speed are due to the tuning parameter q. It turns out that $q > 1$ (along with $\beta > 1$ in the q-DAEM) and $q < 1$ (in the EM-Lq) both help with the convergence speed, even though they have different convergence points. We have proved the first order approximation property in Theorem 4.2.4, which leads to the proof of the monotonicity and the convergence for the EM-Lq. For the q-DAEM, because $q \downarrow 1$ and $\beta \downarrow 1$ make it reduce to the traditional EM, it also converges. When $\beta = 1$ and $q = 1$, both algorithms reduce to the traditional EM algorithm.

4.4 Numerical Results and Validation

Now we compare the performance of two estimators on mixture models: 1) the MLqE from the EM-Lq; 2) the MLE from the EM. We set $q = 0.95$ throughout this section.

4.4.1 Kullback Leibler Distance Comparison

We simulate data using a three component Gaussian mixture model (3GMM)

$$f_0^*(x; \epsilon, \sigma_c^2) = 0.4(1-\epsilon)\varphi(x; 1, 2) + 0.6(1-\epsilon)\varphi(x; 5, 2) + \epsilon\varphi(x; 3, \sigma_c^2). \tag{4.14}$$

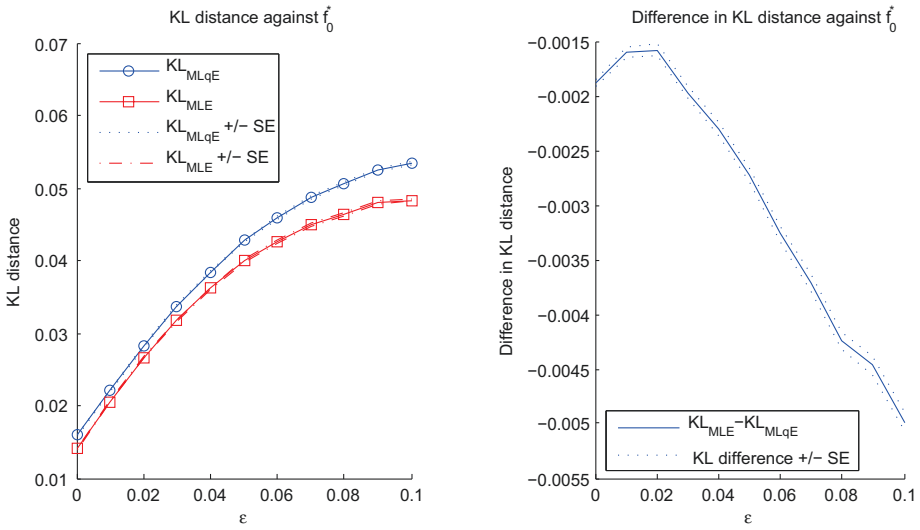
This is a gross error model, where the third term is the outlier component (or contamination component, or measurement error component); ϵ is the contamination ratio ($\epsilon \leq 0.1$); σ_c^2 is the variance of the contamination component, and is usually very large (i.e., $\sigma_c^2 > 10$). Equation (4.14) can be considered as a small deviation from the 2GMM: $f_0(x) = 0.4\varphi(x; 1, 2) + 0.6\varphi(x; 5, 2)$.

As we mentioned in Section 4.1, there are two approaches for estimating f_0 based on data generated by f_0^*. We will investigate them individually.

4.4.1.1 Direct Approach

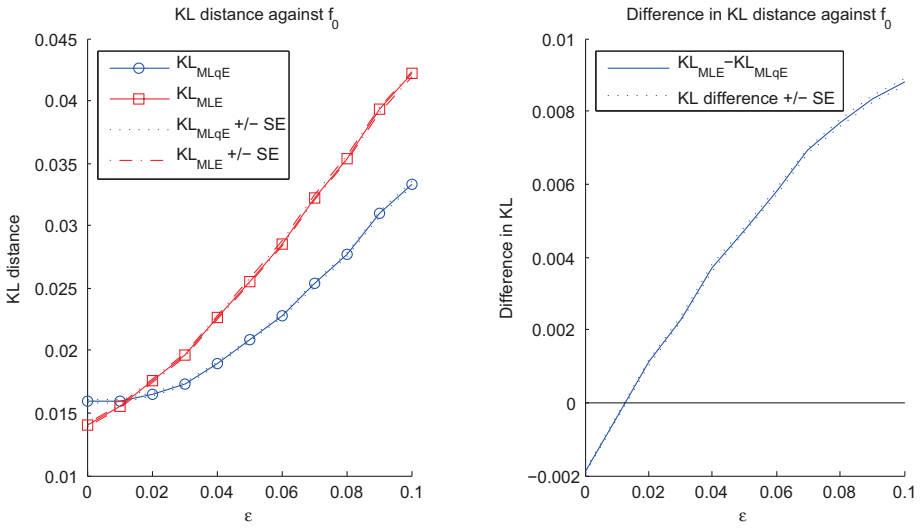
We start with the direct approach. First, we simulate data with sample size $n = 200$ according to equation (4.14), $f_0^*(x; \epsilon, \sigma_c^2 = 20)$, at different contamination levels $\epsilon \in [0, 0.1]$. We fit the 2GMM using the MLqE and the MLE. We repeat this procedure 10,000 times and then calculate (1) the average KL distance between the estimated 2GMM and f_0^*, and (2) the average KL distance between the estimated 2GMM and f_0. We summarize the results in Figure 4.4 (KL against f_0^*) and Figure 4.5 (KL against f_0).

In Figure 4.4a, we see that both $\mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ and $\mathrm{KL}_{\mathrm{MLE}}$ increase as ϵ increases, which means the performance of both MLqE and MLE degrades as more measurement errors are present. $\mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ is always larger than and increases slightly faster than $\mathrm{KL}_{\mathrm{MLE}}$. This implies that the MLqE performs worse than, and degrades faster than the MLE. Figure 4.4b shows their difference $\mathrm{KL}_{\mathrm{MLE}} - \mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ which is negative and decreasing.



(a) (b)

Figure 4.4: Comparison between the MLqE and the MLE in terms of KL distances against f_0^*: (a) shows the KL distances themselves, (b) shows their difference.



(a) (b)

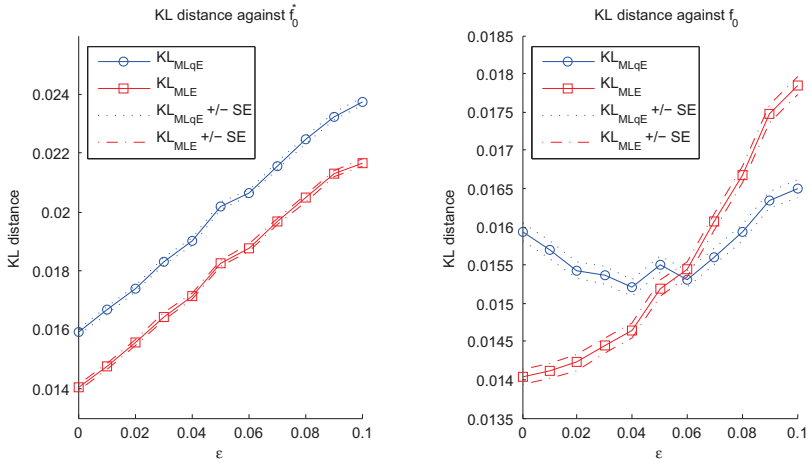
Figure 4.5: Comparison between the MLqE and the MLE in terms of KL distances against f_0: (a) shows the KL distances themselves, (b) shows their difference.

This phenomena is reasonable because, when estimating f_0^* using data generated by f_0^*, the MLE is the best estimator (in terms of KL distance) by definition. The MLqE's bias-variance trade off does not gain anything compared to the MLE.

On the other hand, Figure 4.5 shows an interesting phenomena. In Figure 4.5a, we see that both $\mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ and $\mathrm{KL}_{\mathrm{MLE}}$ still increase as ϵ increases. However, when estimating the non-measurement error components f_0, $\mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ increases more slowly than $\mathrm{KL}_{\mathrm{MLE}}$. The former starts above the latter but eventually ends up below the latter as ϵ increases, which means the MLE degrades faster than the MLqE. Figure 4.5b shows their difference $\mathrm{KL}_{\mathrm{MLE}} - \mathrm{KL}_{\mathrm{ML}q\mathrm{E}}$ which starts negative and increases gradually to positive (changes sign at around $\epsilon = 0.025$). This means that our MLqE performs better than the MLE in terms of estimating f_0 when there are more measurement errors in the data. Hence, we gain robustness from the MLqE.

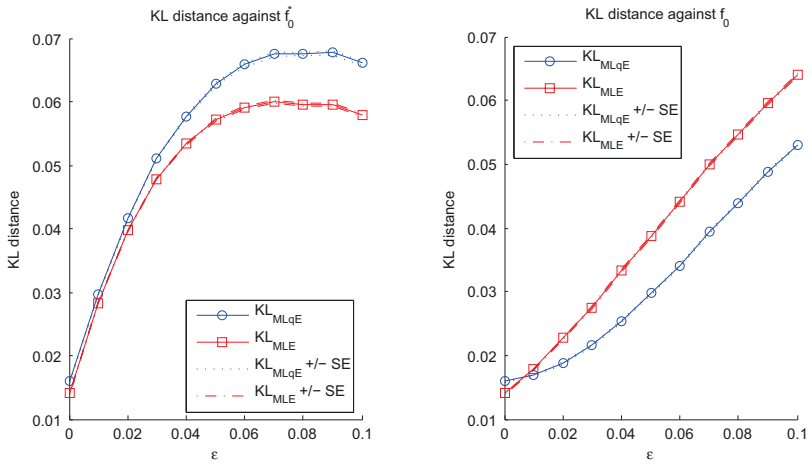
The above simulation is done using the model $f_0^*(x; \epsilon, \sigma_{\mathrm{c}}^2 = 20)$. To illustrate the effect of σ_{c}^2 on the performance of the MLqE, we change the model to $f_0^*(x; \epsilon, \sigma_{\mathrm{c}}^2 = 10)$ and $f_0^*(x; \epsilon, \sigma_{\mathrm{c}}^2 = 30)$, and repeat the above calculations. The results are shown in Figures 4.6 and 4.7.

As we can see, σ_{c}^2 has a big impact on the performance of the estimator. As σ_{c}^2 gets larger (i.e., more serious measurement error problems), both the MLqE and the MLE degrade faster as the contamination ratio increases. This is why the slopes of the KL distance curves become steeper with the higher σ_{c}^2. However, the advantage of the MLqE over the MLE is more obvious with the larger σ_{c}^2. The point where



(a) (b)

Figure 4.6: Comparison between the MLqE and the MLE in terms of KL distances against f_0^* (left panel) and f_0 (right panel) with the third component variance σ_c^2 being 10.



(a) (b)

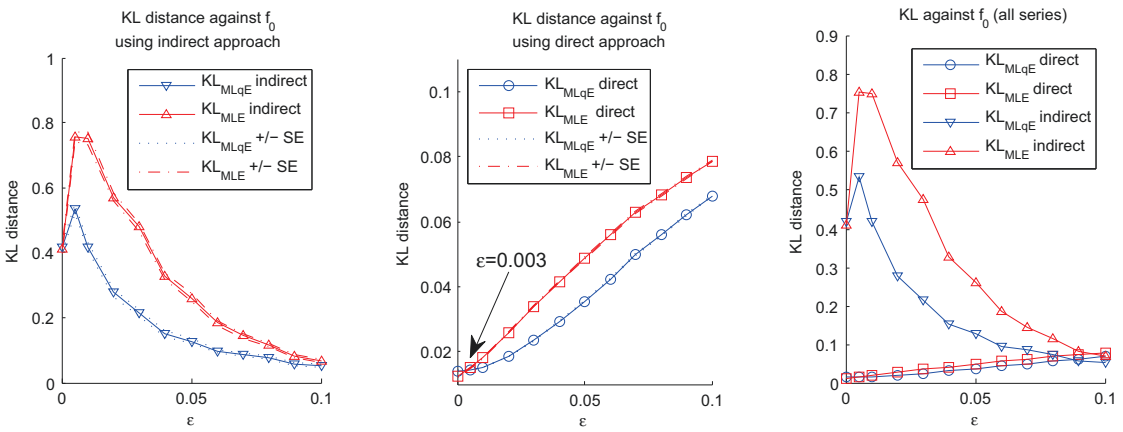
Figure 4.7: Comparison between the MLqE and the MLE in terms of KL distances against f_0^* (left panel) and f_0 (right panel) with the third component variance σ_c^2 being 30.

the two KL distance curves insect (in Figure 4.6b and 4.7b) moves to the left as σ_c^2 increases, which means the MLqE will beat the MLE at a lower contamination ratio in the presence of larger σ_c^2 is used (i.e., the higher variance of the measurement errors).

4.4.1.2 Indirect Approach

Now, let us take the indirect approach, which is to estimate f_0^* first and project it onto the 2GMM space. In this experiment, we let the data be generated by $f_0^*(x; \epsilon, \sigma_c^2 = 40)$ which has an even higher variance of the measurement error component than the previous section. We use a sample size of $n = 200$. We simulate data according to $f_0^*(x; \epsilon, \sigma_c^2 = 40)$, use the ML$q$E and the MLE to fit the 3GMM, take out its component with the largest variance and normalize the weights to get our estimate for f_0. We repeat this procedure 10,000 times, and calculate the average KL distance between our estimates (both the MLqE and the MLE) and f_0. For comparison purposes, we repeat the calculation using the direct approach on this simulation data as well, and summarize the results in Figure 4.8.

In Figure 4.8a, we see that, as ϵ increases, KL distances of the indirect approach first increase and then decrease. The increasing part suggests that a few outliers will hurt the estimation of the non-measurement error component. The decreasing part means that, after the contamination increases beyond a certain level ($\epsilon = 0.5\%$), the more contamination there is, the more accurate our estimates are. This is because that, when the contamination ratio is small, it is hard to estimate the measurement



(a) (b) (c)

Figure 4.8: Comparison between the MLqE and the MLE in terms of KL distances against f_0: (a) shows KL distances obtained from the indirect approach, (b) shows KL distances obtained from the direct approach, (c) shows both these two kinds of KL distances together in order to compare their magnitude.

error component as there are very few outliers. As the contamination ratio gets larger, the indirect approach can more accurately estimate the measurement error component, hence provide better estimates of the non-measurement error components. Please note that our MLqE is still doing better than the MLE in this case. The reason is that the MLqE successfully trades bias for variance to gain in the overall performance. However, as ϵ increases, the advantage of the MLqE gradually disappears. This is because when the contamination is obvious, the MLE will be more powerful and efficient than the MLqE under the correctly specified model.

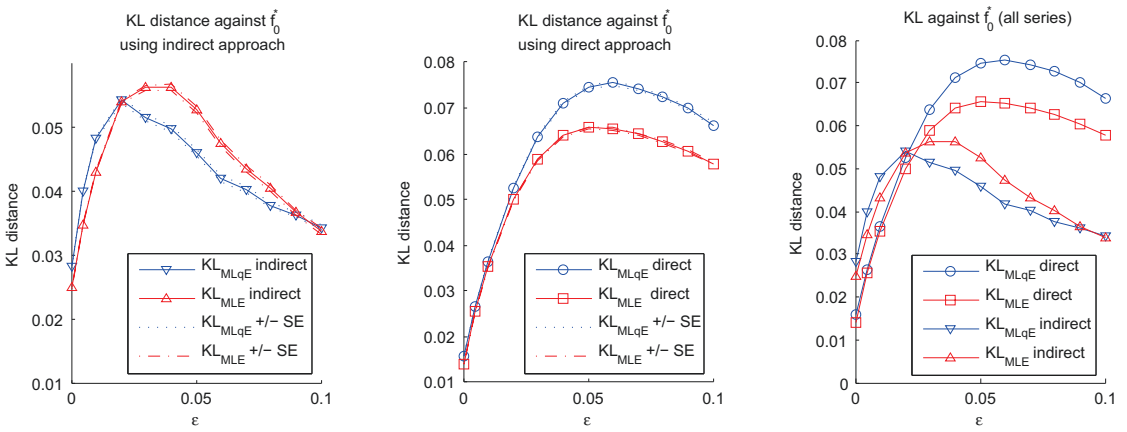
In Figure 4.8b, we present the results for the direct approach, which is consistent with Figure 4.5a. We notice that, when f_0^* has a larger variance for the measure error component, the MLqE beats the MLE at a lower contamination ratio ($\epsilon = 0.003$). In other words, as f_0^* is further deviated from f_0 (in terms of the variance of measurement

error component), the advantage of the MLqE becomes more significant.

In Figure 4.8c, we plot KL distances of both approaches. It is obvious that the indirect approach is much worse than the direct approach until ϵ raises above 0.08. This is because we estimate more parameters and have more estimation variance for the indirect approach. Although our model is correctly specified, the estimation variance is so big that it dominates the overall performance. To summarize, with a small contamination ratio, we are better off using the direct approach with the misspecified model. When the contamination ratio is large, we should use the indirect approach with the correctly specified model.

The above comparison is done based on the KL distance against f_0. We repeat the above calculation to obtain the corresponding results for the KL distance against f_0^*. Note that all the calculations are the same except we do not need to do the projection from 3GMM to 2GMM, because f_0^* is 3GMM. The results are shown in Figure 4.9.

As we can see from Figure 4.9a, when the contamination ratio increases, the KL distances against f_0^* (for both the MLqE and the MLE) increase first and then decrease. This means that as outliers are gradually brought into the data, they first undermine the estimation for the non-measurement error components, and then help the estimation of the measurement error component. The MLqE starts slightly above the MLE. When outliers become helpful for the estimation ($\epsilon > 2\%$), the MLqE goes below the MLE. As ϵ increases beyond 2%, the advantage of the MLqE over the MLE first increases and then diminishes. Figure 4.9b is also consistent with what we



(a) (b) (c)

Figure 4.9: Comparison between the MLqE and the MLE in terms of KL distances against f_0^*: (a) shows KL distances obtained from the indirect approach, (b) shows KL distances obtained from the direct approach, (c) shows both these two kinds of KL distances together in order to compare their magnitude.

found in Figures 4.4a and 4.6a and 4.7a. In Figure 4.9c, we see that the direct and indirect approaches are in about the same range. They intersect at around $\epsilon = 2\%$, which suggests that, when estimating f_0^*, we prefer the direct approach for the mildly contaminated data, and prefer the indirect approach for the heavily contaminated data.

4.4.2 Relative Efficiency

We can also compute the relative efficiency between the MLE and the MLqE using the same model (equation (4.14)), $f_0^*(x; \epsilon, \sigma_c^2 = 20)$.

At each level $\epsilon \in [0, 0.1]$, we generate 3,000 samples with sample size $n = 100$ according to equation (4.14), $f_0^*(x; \epsilon, \sigma_c^2 = 20)$, fit the 2GMM to the data using the

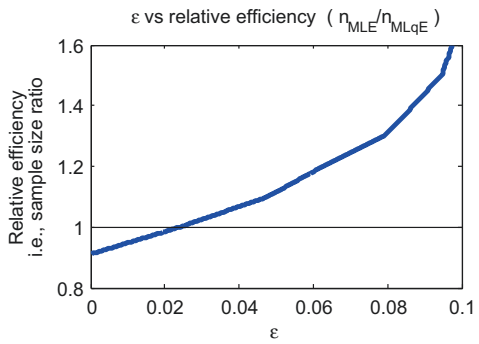


Figure 4.10: Comparison of the MLE and the MLqE based on relative efficiency.

MLqE, and calculate the average KL against f_0. We try the same procedure for the MLE, and find the sample size $n_{\text{MLE}}(\epsilon)$ at which the same average KL is obtained by the MLE. We plot the ratio of these two sample sizes $n_{\text{MLE}}(\epsilon)/100$ in Figure 4.10.

As we can see, the relative efficiency starts below 1, which means, when the contamination ratio is small, it takes the MLE fewer samples than the MLqE to achieve the same performance. However, as the contamination ratio increases, the relative efficiency climbs substantially above 1, meaning that the MLE will need much more data than the MLqE to achieve the same performance.

4.4.3 Gamma Chi-Square Mixture Model

We take a small digression and consider estimating a Gamma Chi-square mixture model,

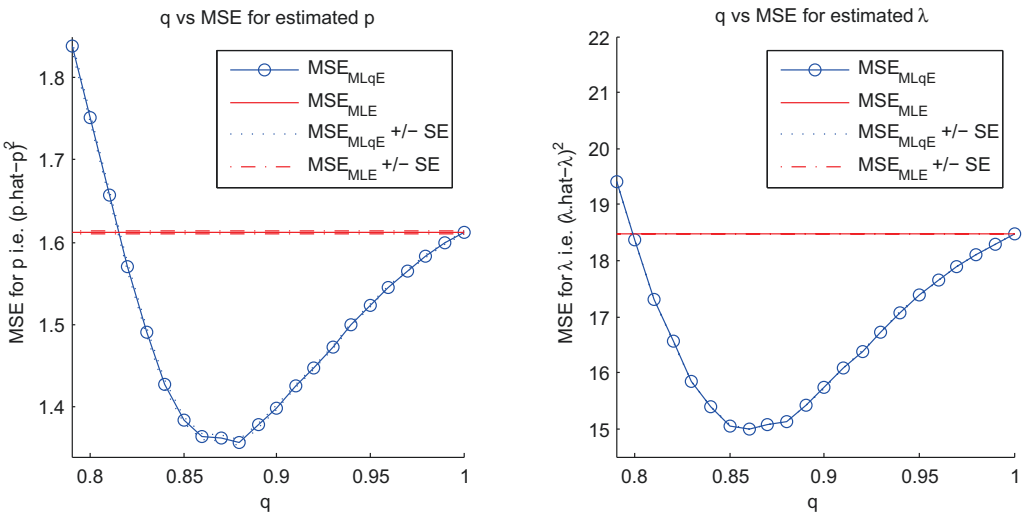
$$f_0^*(x) = (1-\epsilon)\text{Gamma}(x; p, \lambda) + \epsilon\chi^2(x; d), \tag{4.15}$$

where the second component is the measurement error component. We can think of our data being generated from the Gamma distribution but contaminated with the Chi-square gross error. In this section, we consider two scenarios:

Scenario 1: $p = 2$, $\lambda = 5$, $d = 5$, $\epsilon = 0.2$, $n = 20$

Scenario 2: $p = 2$, $\lambda = 0.5$, $d = 5$, $\epsilon = 0.2$, $n = 20$

In each scenario, we generate 50,000 samples according to equation (4.15), fit the Gamma distribution using both the MLqE and the MLE, and compare these two estimators based on their mean square error (MSE) for p and λ. For the MLqE, we adjust q to examine the effect of the bias-variance trade off. The results are summarized in Figure 4.11 (scenario 1) and Figure 4.12 (scenario 2). In Figure 4.11, We see that, by setting $q < 1$, we can successfully trade bias for variance and obtain better estimation. In scenario 1, since the Gamma distribution and the Chi-square distribution are sharply different, the bias-variance trade off leads to a significant reduction on the mean square error by partially ignoring the outliers. However, in scenario 2, these two distributions are similar (the mean and variance of the Gamma distribution are 4 and 8, the mean and variance of the Chi-square distribution are 5 and 10). In this situation, partially ignoring the data points on the tails will not help much, which is why the MSE of the MLqE is always larger than the MSE of the MLE.



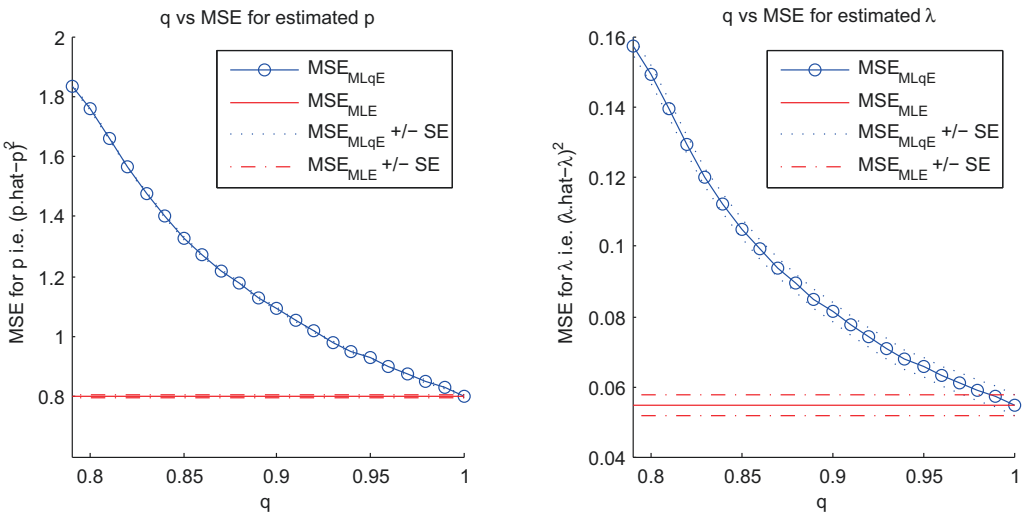
(a) (b)

Figure 4.11: Comparison of the MLE and the MLqE in terms of the MSE for $\hat{p}$ (Figure a) and $\hat{\lambda}$ (Figure b) in scenario 1 ($p = 2$, $\lambda = 5$, $d = 5$, $\epsilon = 0.2$, $n = 20$).

4.4.4 Old Faithful Geyser Eruption Data

We consider the Old Faithful geyser eruption data from [20]. The original data is obtained from the R package "tclust". The data is univariate eruption time length with sample size of 272. We sort these eruption lengths by their times of occurrences, and lag these lengths by one occurrence to form 271 pairs; thus we have two dimensional data (i.e., current eruption length and previous eruption length). This is the same procedure as described in [21]. For this two dimensional data, they have suggested three clusters. Since the "short followed by short" eruptions are not usual, [21] identify these points in the lower left corner as outliers.

We plot the original data in Figure 4.13, fit the MLqE ($q = 0.8$) and the MLE to the data, and plot the 2 standard deviation ellipsoids. q is selected based on



(a) (b)

Figure 4.12: Comparison of the MLE and the MLqE in terms of the MSE for $\hat{p}$ (Figure a) and $\hat{\lambda}$ (Figure b) in scenario 2 ($p = 2$, $\lambda = 0.5$, $d = 5$, $\epsilon = 0.2$, $n = 20$).

clustering outcome. As we can see, there are a few outliers in the lower left corner. The MLE is obviously affected by the outliers. The lower right component of the MLE is dragged to the left to accommodate these outliers, and thus misses the center of the cluster. Other components of the MLE are also mildly affected. The MLqE, on the other hand, overcomes this difficulty and correctly identifies the center of each component. This improvement is especially obvious for the lower right component: the fitted MLqE lies in the center whereas the MLE is shifted to the left and has a larger 2 standard deviation ellipsoid.

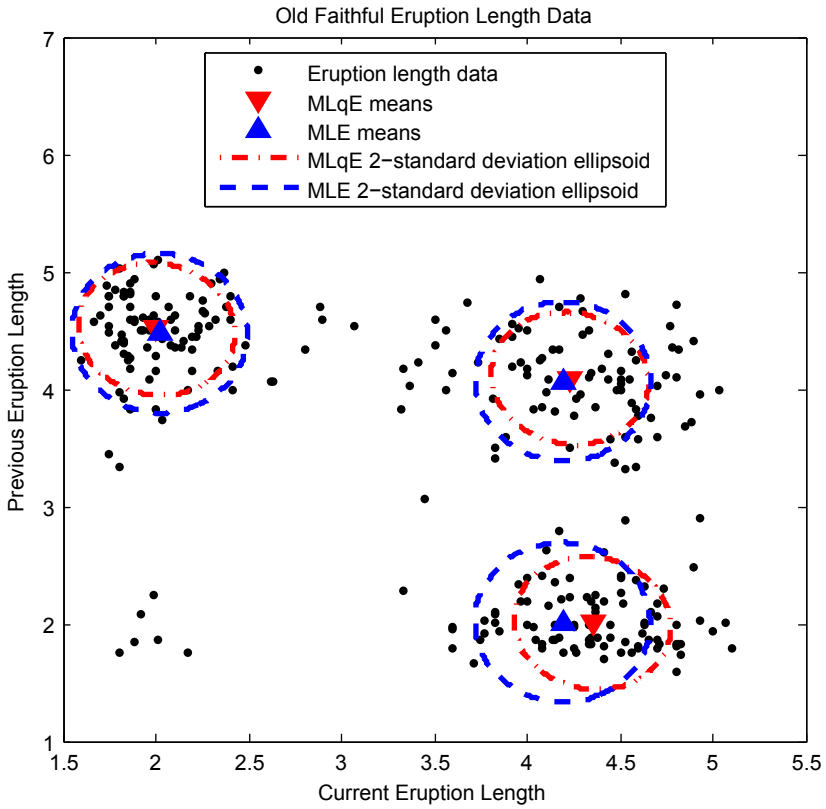


Figure 4.13: Comparison between the MLqE and the MLE for the Old Faithful geyser data: red triangles: MLqE means; red dashed lines: MLqE two standard deviation ellipsoids; blue triangles: MLE means; blue dashed lines: MLE two standard deviation ellipsoids.

4.5 Selection of q

So far in this chapter, we have fixed q in all the analysis. In this section, we will investigate the selection of q.

The tuning parameter q governs the sensitivity of the estimator against outliers. The smaller q is, the less sensitive the MLqE is to outliers. If the contamination becomes more serious (i.e., larger ϵ and/or σ_c^2), we should use a smaller q to protect against measurement errors. There is no analytical relation between the level of contamination and q, because it depends on the properties of the non-measurement error components, the contamination ratio and the variance of the contamination component. Furthermore, there is no guarantee that the measurement error component is a normal distribution.

Generally, it is very hard to choose q analytically. Currently, there is no universal way to do so. Instead, we here present an example to illustrate the idea of selecting q. We generate one data set using equation (4.14) $f_0^*(x; \epsilon = 0.1, \sigma_c^2 = 40)$ with the sample size $n = 200$. We will demonstrate how to select q for this particular data set.

First, we fit a 3GMM to the data using the MLE and get $\hat{f}_{3\text{GMM}}$. We identify the component with the largest variance in $\hat{f}_{3\text{GMM}}$ as the contamination component. We extract the *non-measurement error components* and renormalize weights to get $\hat{f}_{3\text{GMM}\to 2\text{GMM}}$, which can be considered as the projection from the 3GMM to the 2GMM space. We go back to $\hat{f}_{3\text{GMM}}$, utilize it to perform a parametric bootstrap by generating many bootstrap samples, and fit 2GMM to these data sets using MLqE

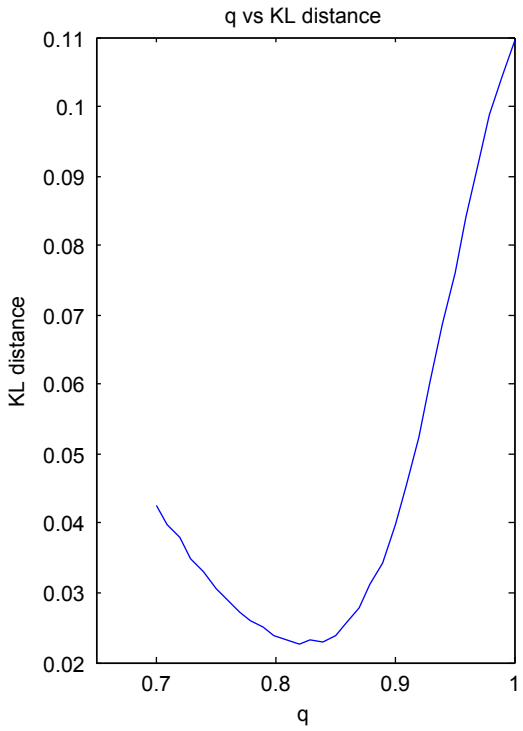


Figure 4.14: Selection of q based on average KL distance from the bootstrap samples.

$(\hat{f}^{b\,\mathrm{ML}q\mathrm{E}}_{2\mathrm{GMM}})$ with q varying between 0.7 and 1. We take the q that minimizes the average KL distance between the *non-measurement error components*, $\hat{f}_{3\mathrm{GMM}\to 2\mathrm{GMM}}$, and the estimated 2GMM $\hat{f}^{b\,\mathrm{ML}q\mathrm{E}}_{2\mathrm{GMM}}$ from the bootstrap samples. The average KL distance against q is shown in Figure 4.14. From the figure, we estimate q_{optimal} to be 0.82.

This is a very simple way to select q. It is straightforward and easy. However, there is a drawback of this method. When the contamination ratio is very low (e.g., 1% or 2%) and the sample size is small ($n < 100$), the estimated 3GMM $\hat{f}_{3\mathrm{GMM}}$ will not be able to estimate the measurement error component correctly since there are very few outliers. Thus, the parametric bootstrap approach following that will

become unreliable. We have not found an effective way of selecting q with the small contamination ratio.

In [1], they have mentioned using asymptotic variance and asymptotic efficiency as criteria for selecting q. However, obtaining the variance in the mixture model case is also problematic and unreliable when sample size is small.

To obtain an analytical solution for q is hard. Currently, we have only some remedies under a few situations, and are still looking for a universal way. However, we believe that selecting q is a very important question and is one of major future research directions.

4.6 Conclusion

In this chapter, we have introduced a new estimation procedure for mixture models, namely the MLqE, along with the EM-Lq algorithm. Our new algorithm provides a more robust estimation for mixture models when measurement errors are present in the data. Simulation results show superior performance of the MLqE over the MLE in terms of estimating the non-measurement error components. Relative efficiency is also studied and shows superiority of the MLqE. Note that when $q = 1$, the MLqE becomes the MLE, so the MLqE can be considered as a generalization of the MLE.

Throughout this chapter, we see that the MLqE works well with mixture models in the EM framework. There is a fundamental reason for such a phenomenon. Note

that the M step of the traditional EM solves a set of weighted likelihood equations with weights being the soft labels. Meanwhile, the MLqE solves a different set of weighted likelihood equations with weights being f^{1-q}. Therefore, incorporating the MLqE in the EM framework comes down to determining the new weights that are consistent with both the soft labels and f^{1-q}. Furthermore, we conjecture that, for any new types of estimators, as long as they involve only solving sets of weighted likelihood equations, they should be able to be smoothly incorporated in the mixture model estimation using the EM framework.

In order to achieve consistency for the MLqE, we need the distortion parameter q to approach 1 as the sample size n goes to infinity. However, letting q converge to 1 will affect the bias-variance trade off. So what is the optimal rate at which q tends to 1 as $n \to \infty$? Meanwhile, how to select q at different sample sizes is also an interesting topic. The distortion parameter q adjusts how aggressive or conservative we are towards eliminating the effect of outliers. Tuning of the distortion parameter q will be a fruitful direction for future research.

Chapter 5

An Application to Brain Graph Data

In this chapter, we introduce a robust clustering technique for analyzing brain graph data using Maximum Lq-likelihood Estimation (MLqE). The methodology is based on the EM-Lq algorithm previously proposed in Chapter 4 and [22]. We present a comparison of MLE and MLqE in terms of Adjusted Rand Index (ARI) and demonstrate the superior performance of MLqE.

5.1 Description of Data

We receive the adjacency spectral embeddings of one brain graph obtained by the methodology proposed in [23]. The original brain graph data is generated using the

pipeline described in [24]. The brain graph is divided into $R = 70$ non-overlapping regions where each vertex belongs to only one region. The data takes the form of matrix $X = (x_{up})^{n \times P}$. The sample size is $n = 543742$ (i.e., 543742 vertices in the brain graph). Dimension size is $P = 50$ (i.e., embeddings in a 50 dimensional space). $\vec{x}_u = (x_{u1}, ..., x_{uP})$ is a vector with length of $P = 50$, which represents the embedding of the vertex u from the brain graph into the $P = 50$ dimensional space. In addition, another vector $Y = (y_1, ..., y_n)$ of length $n = 543742$ is provided which contains the true region label of each vertex. We have in total $R = 70$ regions, so $y_i \in \{1, 2, ..., 70\}$.

5.2 Methodology

In this analysis, we use Gaussian mixture models to fit the data using Maximum Likelihood Estimation and Maximum Lq-Likelihood Estimation. The MLE is obtained by EM algorithm [25] which is implemented in `mclust()`. The MLqE is obtained by EM-Lq algorithm which is implemented in `qclust()`. When $q = 1$, MLqE becomes MLE, and `qclust()` is equivalent to `mclust()`.

For each region pair $\{(i, j)|i, j = 1, ..., 70, i < j\}$ among $\binom{70}{2}$ region pairs, we extract the rows from X which belong to the two regions (say, region i and j), and further take a subsample with sample size $m = 800$ (400 cases randomly selected from region i and another 400 cases randomly selected from region j). After deleting the $r + 1$st dimension through the 50th dimension, we end up with a data matrix $\tilde{X}_{ij}$ (m

by r) and $\tilde{Y}_{ij}$ (m by 1) where each element of $\tilde{Y}_{ij}$ is either i or j.

We fit a Gaussian mixture model (GMM) to the data $\tilde{X}_{ij}$ with `mclust()` using "VVV" model type. We further fit a GMM to $\tilde{X}_{ij}$ with `qclust()` (the complexity of the GMM is provided by `mclust()`, the initial parameters for `qclust()` are also given by the estimate obtained from `mclust()`). We set $q = 0.98$ in the entire analysis for demonstration.

Based on the GMMs estimated by `mclust()` and `qclust()`, we will be able to assign cluster memberships to all data points denoted as $Z_{ij}^{(1)} = (z_{ij,1}^{(1)}, ..., z_{ij,m}^{(1)})$ and $Z_{ij}^{(q)} = (z_{ij,1}^{(q)}, ..., z_{ij,m}^{(q)})$. $Z_{ij}^{(1)}$ and $Z_{ij}^{(q)}$ are two vectors with length of $m = 800$. Notice that $z_{ij,i}^{(1)}, z_{ij,i}^{(q)} \in \{1, 2, ..., \hat{K}_{ij}\}$ where $\hat{K}_{ij}$ is the complexity of GMM estimated by `mclust()` for $\tilde{X}_{ij}$. We further calculate the ARIs of these two methods (i.e., $\text{ARI}_{ij}^{(1)}$ and $\text{ARI}_{ij}^{(q)}$). $\text{ARI}_{ij}^{(1)} = \text{adjustedRandIndex}(Z_{ij}^{(1)}, \tilde{Y}_{ij})$ and $\text{ARI}_{ij}^{(q)} = \text{adjustedRandIndex}(Z_{ij}^{(q)}, \tilde{Y}_{ij})$. We calculate the difference between the two ARIs, $d_{ij} = \text{ARI}_{ij}^{(q)} - \text{ARI}_{ij}^{(1)}$.

We repeat the above calculation for all the region pairs $\{(i, j)|i, j \in \{1, ..., 70\}, i < j\}$ and collect all d_{ij}s. We conduct the Wilcoxon test on the hypothesis:

$$H_0 : \text{median}(d_{ij}) = 0$$

$$H_A : \text{median}(d_{ij}) > 0$$

$\hat{K}_{ij}$	frequency	$\#(d_{ij} > 0)$	$\#(d_{ij} < 0)$	Wilcoxon p-value
1	1897	0	0	NA
2	310	194	111	0.0000
3	92	46	45	0.4170
4	56	28	26	0.3417
5	30	10	20	0.9506
6	13	3	9	0.9270
7	11	4	5	0.5000
8	3	2	0	0.0000
9	3	2	1	0.1250
Total		**289**	**217**	**0.0006**

Table 5.1: Summary of Wilcoxon tests for different $\hat{K}_{ij}$ using the first two dimensions ($r = 2$)

5.3 Results

5.3.1 Wilcoxon Tests

The results of Wilcoxon tests are summarized in Table 5.1 and Table 5.2. Table 5.1 shows the Wilcoxon tests using the first $r = 2$ dimensions of X, while Table 5.2 shows the Wilcoxon tests using the first $r = 4$ dimensions of X. We also show the results of Wilcoxon tests conditioned on $\hat{K}_{ij}$.

From these tables, we see that when using the first 2 dimensions of X, the overall Wilcoxon test (in bold numbers) rejects the null hypothesis that median$(d_{ij}) = 0$. When conditioning on $\hat{K}$, we see that such Wilcoxon tests are rejected when $\hat{K}_{ij} = 2$ and are not rejected otherwise. Moreover, when using the first 4 dimensions of X, the overall Wilcoxon test (in bold numbers) rejects the null hypothesis with a more significant p-value. When conditioned on $\hat{K}_{ij}$, Wilcoxon tests at $\hat{K}_{ij} = 2, 3, 4, 5, 9$ are

$\hat{K}_{ij}$	frequency	$\#(d_{ij} > 0)$	$\#(d_{ij} < 0)$	Wilcoxon p-value
1	1825	0	0	NA
2	332	193	130	0.0002
3	109	66	34	0.0004
4	43	28	14	0.0098
5	37	24	13	0.0235
6	33	19	14	0.1481
7	19	11	7	0.1189
8	8	5	3	0.1445
9	9	7	2	0.0195
Total		**353**	**217**	**0.0000**

Table 5.2: Summary of Wilcoxon tests for different $\hat{K}_{ij}$ using the first four dimensions ($r = 4$)

all rejected.

To summarize, we understand that MLqE provides a better clustering result than MLE because it partially ignores the outliers which are very common in the data set X. This phenomenon becomes more significant when applied in higher dimensional space ($r = 4$ compared to $r = 2$).

Since we are using $\binom{70}{2}$ region pairs, our d_{ij}s are correlated which undermines the p-values obtained by Wilcoxon tests. However, the correlation is reduced when we take subsamples (sample size $m = 800$) from the original data set X.

5.3.2 A In-Depth Example of One Region Pair

In this section, we take a particular pair: region 8 and 38, and present the result associated with this pair.

We first plot the first two dimensions of the data from this region pair in Figure

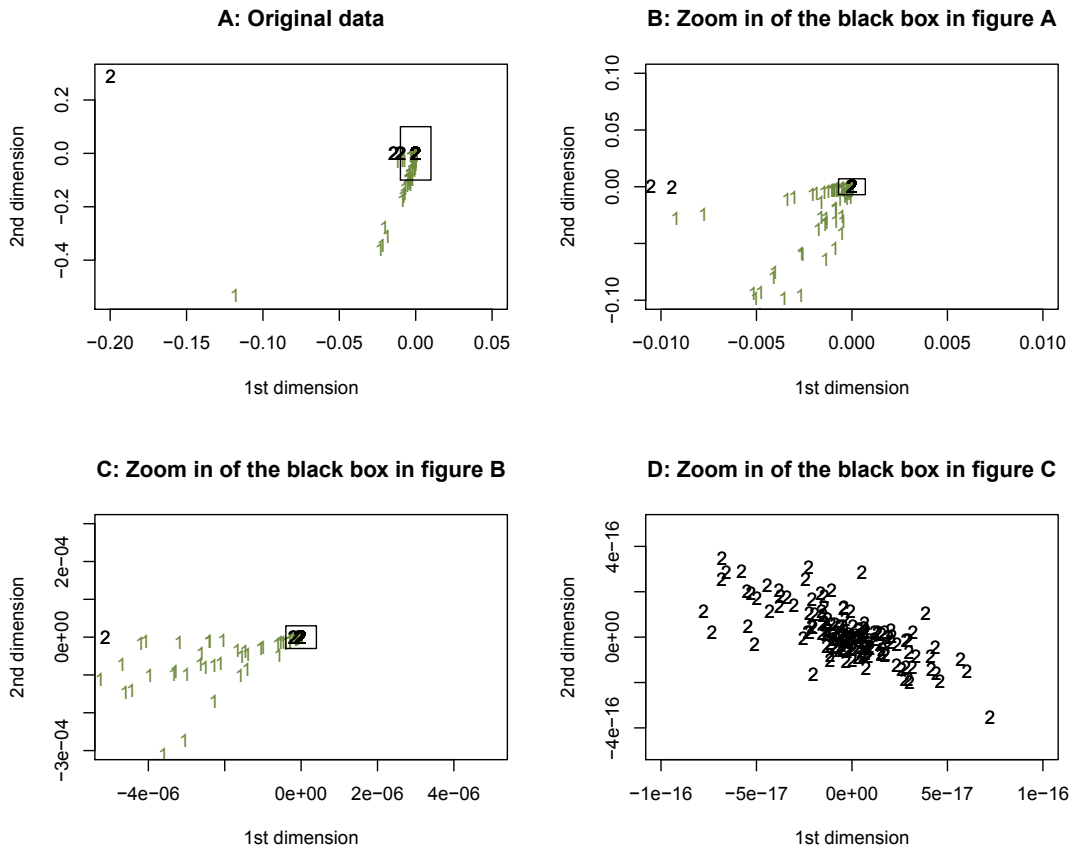


Figure 5.1: The embedding data of region pair 8 and 38. Panel A displays the orginal data at the orignal scale, black "2"s— region 8, green "1"s — region 38; Panel B displays the black box in panel A at a smaller scale; Panel C displays the black box in panel B at an even smaller scale; Panel D displays the black box in panel C at the smallest scale.

5.1. In panel A, we clearly see there are many apparent outliers. It also appears that many data points (especially black "2"s) are concentrated in a very small area. Therefore, in panels B, C and D, we gradually zoom in around the center of the black "2"s and finally see the shape of this cluster.

We further fit GMMs to the data using MLE and MLqE. The clustering results are shown in Figure 5.2. The red and blue curves (one standard deviation) indicate

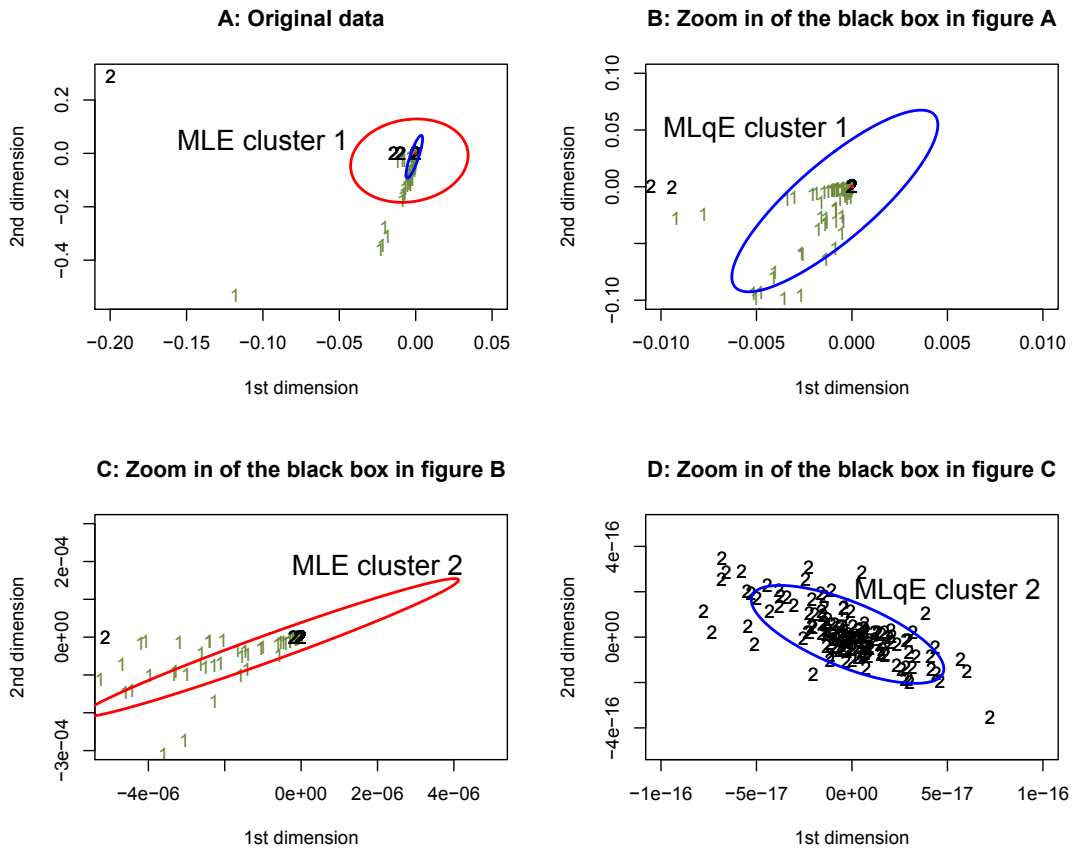


Figure 5.2: The clustering results for region pair 8 and 38. Red curve: one standard deviation ellipsoid of the normal distributions of each cluster fitted by MLE; Blue curve: one standard deviation ellipsoid of the normal distributions of each cluster fitted by MLqE. Panels A, B, C and D still display the same regions and same scales as in Figure 5.1

the clustering given by MLE and MLqE. From panel A, we can see the first MLE cluster is obviously affected by the outliers, whereas the first MLqE cluster is much more accurate and captures the shape of green "1"s. The same idea appears in panel B. In panel C, we see the second MLE cluster is still identifying green "1"s. On the other hand, in panel D the second MLqE cluster correctly identifies the shape and location of the center of black "2"s.

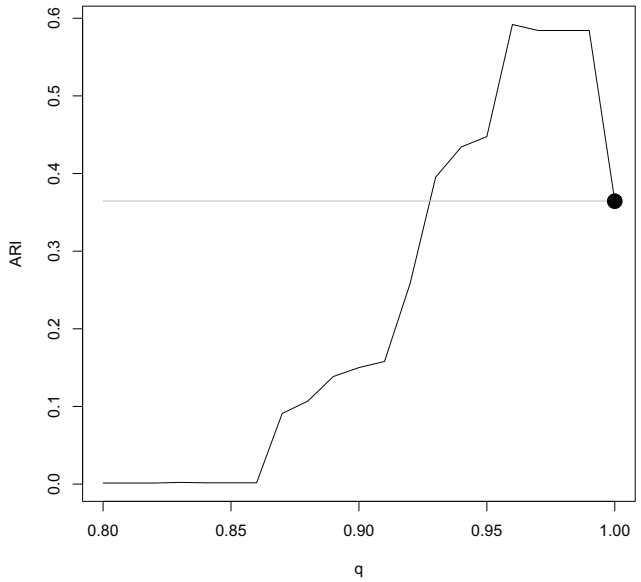


Figure 5.3: The change of ARI as q goes from 1 to 0.8 for region pair 8 and 38.

We further plot $\text{ARI}_{8,38}(q)$ as a function of q, the tuning parameter in MLqE, in Figure 5.3. As we can see from the figure, by lowering q away 1, we get significant improvement in terms of ARI first and then the advantage gradually disappears as q keeps moving away 1. As q reaches 0.85, we have ARI approach 0. Notice that when $q = 0.96$, we have the highest ARI which is about 50% higher than the ARI at $q = 1$ (the black dot) which corresponds to MLE (i.e., `mclust()`).

From Figure 5.1 and Figure 5.2, we know that, because `mclust()` gives each data point equal weight, the two Gaussians estimated from `mclust()` are largely affected by outliers. On the other hand, the GMM given by `qclust()` is correctly identifying the centers of the blue dots and red dots by ignoring outliers (especially these outliers that are very distant from the centers which have huge effect on the estimation of

`mclust()`).

5.4 Conclusion

In this chapter, we have applied the MLqE of Gaussian mixture models on the adjacency embeddings of the brain graph data and obtained superior performance compared to the results obtained by MLE. By partially ignoring outliers in the data, MLqE can provide a better clustering results than MLE in terms of ARI.

Chapter 6

Discussion

In this chapter, we conclude our current research as well as present future research directions.

6.1 Conclusion

Robust statistics is one of the most important research areas in statistics because real world problems seldom fit these strict assumptions perfectly. Model assumption violation or model misspecification are ubiquitous. The Lq-likelihood approach we present in this dissertation functions as a special class of Huber's M-estimators which includes the traditional maximum likelihood estimator as a special case. The whole class of estimators is indexed by a tuning parameter q which reduces the original class of M-estimator (indexed by ψ) to a one dimensional space. More importantly,

this class of M-estimator gives the solution of the weighted likelihood equation with special form of weights. Another advantage of the MLqE is that sometimes Huber's M-estimator's ψ function is difficult to interpret; there may not even exist a contrast function ρ corresponding to it. On the other hand, the MLqE has a very interpretable contrast function — Lq-likelihood. Moreover, this Lq-likelihood gives nice properties in other areas of statistics, for example, Bayesian statistics. The delicate design of the Lq-likelihood connects Huber's ψ function with a well defined contrast function ρ which leads to many desirable properties of the estimator.

The Lq-likelihood effectively brings robustness to traditional statistical inference while maintaining small efficiency loss. The limiting case of Lq-likelihood as $q \to 1$ means that we can use Lq-likelihood as a framework to trade off between efficiency and robustness.

6.2 Future Research

For future research, we will focus on several topics.

First of all, the selection of the tuning parameter remains a challenging issue. For estimation purposes, the selection of q is essentially the problem of allocating the weights so that the estimate is the most accurate. On the other hand, for testing purposes, the selection of q becomes the problem of maximizing the power function. It is reasonable to develop different procedures for the selection of q for estimation

and testing. No matter whether estimation or testing, selecting q is essentially the problem of efficiency and robustness trade off. Such a trade off may appear easy under certain situations while extremely hard under others. Therefore, selecting q needs to be addressed within the context specified.

We understand that solving MLqE is equivalent to solving the weighted likelihood equation. Therefore, the connection between MLqE and other weighted likelihood equation approaches is worth investigating. The ways to decide weight allocation in other approaches may also shed some light on how to select q.

We have been discussing the problem of statistical inference under model misspecification or assumption violations. However, how to detect model misspecification is another interesting problem. In this case, we suspect we will need $q > 1$ to "exaggerate" the effect. Proposing a model misspecification test is promising and seems to be a fruitful direction for future research as well.

Chapter 7

Appendix

7.1 Assumptions 1 - 4

Assumptions 1 - 4 of [12] pp 371-372 are restated:

1) The function u_q is differentiable at θ_0 with the derivative $u'_q(\theta_0) \neq 0$;

2) The standard deviation of $T_{q,n}$ is of order $1/\sqrt{n}$;

3) For a sequence of alternative $\theta_n \to \theta_0$, the distribution of $[T_{q,n} - u_q(\theta_n)]/\sqrt{V_q(\theta_n)/n}$ tends to the standard normal distribution, where $\theta_n \to \theta_0$ as $n \to \infty$;

4) $V_q(\theta_n)/V_q(\theta_0) \to 1$.

7.2 Proof of Theorem 3.3.3

First, we know that $B(\epsilon, q=1) = -E_h\left[\frac{f''_\theta}{f}\right] + A(\epsilon, q=1) = -\epsilon E_g\left[\frac{f''_\theta}{f}\right] + A(\epsilon, q=1)$, where we use the fact that $E_f[f''_\theta/f] = 0$ in the last step. Since $E_g[f''_\theta/f] > 0$, we have $A(\epsilon, q=1) > B(\epsilon, q=1) > 0$, and hence, $\frac{A(\epsilon,q=1)}{B(\epsilon,q=1)} > 1$. When f is a normal distribution, $E_g[f''_\theta/f] > 0$ becomes $0 < E_g\left[\frac{f''_\theta}{f}\right] = E_g\left[\frac{(x-\theta)^2}{\sigma^4} - \frac{1}{\sigma^2}\right] = \frac{\sigma_g^2}{\sigma_f^4} - \frac{1}{\sigma_f^2}$, which is equivalent to $\sigma_f^2 < \sigma_g^2$. Note that this condition does not require g to be a normal distribution.

7.3 Proof of Theorem 3.3.4

Since $h = (1-\epsilon)f + \epsilon g$, we have

$$\frac{\partial}{\partial\epsilon}\Big(A(\epsilon, q=1) - B(\epsilon, q=1)\Big) =$$
$$E_g[\psi_q(X;\theta)^2] + E_g[\psi'_q(X;\theta)] - \Big(E_f[\psi_q(X;\theta)^2] + E_f[\psi'_q(X;\theta)]\Big),$$

which is a function that does not involve ϵ, that is, a constant function in ϵ. Furthermore, we know $A(\epsilon=0, q=1) - B(\epsilon=0, q=1) = 0$ (from Definition 3.3.1), and $A(\epsilon, q=1) - B(\epsilon, q=1) > 0$ for $\epsilon > 0$ (from Theorem 3.3.3). Therefore, we conclude that $\frac{\partial}{\partial\epsilon}(A-B)|_{q=1} > 0$.

7.4 Proof of Theorem 3.3.5

We assume that for a gross error model $h = (1-\epsilon)f + \epsilon g$, we have

$$E_h\Big[\Big[\frac{f''_\theta}{f} + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] < E_f\Big[\Big[\frac{f''_\theta}{f} + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] < 0. \quad (7.1)$$

When both f and g are normal, the condition becomes $\sigma_g^2 > \sigma_f^2(-\frac{2}{3}\log\sqrt{2\pi}\sigma_f - \frac{1}{2})$.

We calculate

$$\begin{aligned}\frac{\partial}{\partial q}\frac{\partial}{\partial \epsilon}(A-B) =& E_g\Big[\Big[(q-2)\Big(\frac{f'_\theta}{f}\Big)^2 - \frac{f''_\theta}{f}\Big]\log f\cdot f^{1-q} - \Big(\frac{f'_\theta}{f}\Big)^2 f^{1-q}\Big] \\ & - E_f\Big[\Big[(q-2)\Big(\frac{f'_\theta}{f}\Big)^2 - \frac{f''_\theta}{f}\Big]\log f\cdot f^{1-q} - \Big(\frac{f'_\theta}{f}\Big)^2 f^{1-q}\Big]\end{aligned}$$

By setting $q=1$, we have

$$\begin{aligned}&\frac{\partial}{\partial q}\frac{\partial}{\partial \epsilon}(A-B)\Big|_{q=1} = \\ &\quad E_f\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] - E_g\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\end{aligned}$$

Since we know

$$\begin{aligned}&E_h\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] < E_f\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] \\ \Rightarrow\ & E_g\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big] < E_f\Big[\Big[\Big(\frac{f'_\theta}{f}\Big)^2 + \frac{f''_\theta}{f}\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\end{aligned}$$

Therefore, we know $\frac{\partial}{\partial q}\frac{\partial}{\partial \epsilon}(A-B)|_{q=1} > 0$. Since $A-B$ is a continuous function of ϵ and q, there exists a C such that $\frac{\partial}{\partial q}\frac{\partial}{\partial \epsilon}|A-B| > 0$ for $C < q \leq 1$. Now let us look at the assumption

$$E_h\Big[\Big[\frac{f''_\theta}{f}+\Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f+\Big(\frac{f'_\theta}{f}\Big)^2\Big] < E_f\Big[\Big[\frac{f''_\theta}{f}+\Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f+\Big(\frac{f'_\theta}{f}\Big)^2\Big] < 0. \quad (7.2)$$

Since $h=(1-\epsilon)f+\epsilon g$, the assumption is equivalent to

$$E_g\Big[\Big[\frac{f''_\theta}{f}+\Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f+\Big(\frac{f'_\theta}{f}\Big)^2\Big] < E_f\Big[\Big[\frac{f''_\theta}{f}+\Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f+\Big(\frac{f'_\theta}{f}\Big)^2\Big] < 0. \quad (7.3)$$

When f is a normal distribution, we have $\frac{f'_\theta}{f}=\frac{x-\theta}{\sigma_f^2}$, $\frac{f''_\theta}{f}=\frac{(x-\theta)^2}{\sigma_f^4}-\frac{1}{\sigma_f^2}$, $\log f = -\log\sqrt{2\pi}\sigma_f-\frac{(x-\theta)^2}{2\sigma_f^2}$. Hence, the assumption becomes

$$\frac{\sigma_g^2}{\sigma_f^4}(-\log 2\pi\sigma_f^2+\frac{3}{2})-K_g\frac{\sigma_g^4}{\sigma_f^6} < \frac{\sigma_f^2}{\sigma_f^4}(-\log 2\pi\sigma_f^2+\frac{3}{2})-K_f\frac{\sigma_f^4}{\sigma_f^6}, \quad (7.4)$$

where K_g and K_f are Kurtosises of g and f. When we assume normal distributions, $K_g=K_f=3$, the assumption further reduces to $\sigma_g^2 > \sigma_f^2(-\frac{2}{3}\log\sqrt{2\pi}\sigma_f-\frac{1}{2})$. Apparently, this assumption holds for most of the cases. Meanwhile, we also require

$$E_f\Big[\Big[\frac{f''_\theta}{f}+\Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f+\Big(\frac{f'_\theta}{f}\Big)^2\Big] < 0.$$

Under the normal distribution assumption, it becomes

$$0 > \frac{\sigma_f^2}{\sigma_f^4}(-\log 2\pi\sigma_f^2 + \frac{3}{2}) + \frac{\log\sqrt{2\pi}\sigma_f}{\sigma_f^2} - K_f\frac{\sigma_f^4}{\sigma_f^6} \Rightarrow \sigma_f > \frac{1}{\sqrt{2\pi}}\exp\{-\frac{3}{2}\}.$$

7.5 Proof of Theorem 3.3.6

We have

$$\begin{aligned}
\frac{\partial}{\partial q}A(\epsilon,q) =& \frac{\partial}{\partial q}E_h\Big[\Big(\frac{f'_\theta}{f}f^{1-q}\Big)^2\Big] = E_h\Big[-2\Big(\frac{f'_\theta}{f}f^{1-q}\Big)^2\log f\Big],\\
\frac{\partial}{\partial q}B(\epsilon,q) =& E_h\Big[\frac{f''_\theta}{f}f^{1-q}\log f + \Big(\frac{f'_\theta}{f}\Big)^2 f^{1-q} - q\Big(\frac{f'_\theta}{f}\Big)^2 f^{1-q}\log f\Big].
\end{aligned}$$

When $q = 1$, we have

$$\begin{aligned}
\frac{\partial}{\partial q}A(\epsilon,q)\Big|_{q=1} =& E_h\Big[-2\Big(\frac{f'_\theta}{f}\Big)^2\log f\Big],\\
\frac{\partial}{\partial q}B(\epsilon,q)\Big|_{q=1} =& E_h\Big[\Big[\frac{f''_\theta}{f} - \Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\\
\frac{\partial}{\partial q}A(\epsilon,q) - B(\epsilon,q)\Big|_{q=1} =& E_h\Big[-\Big[\frac{f''_\theta}{f} + \Big(\frac{f'_\theta}{f}\Big)^2\Big]\log f - \Big(\frac{f'_\theta}{f}\Big)^2\Big] > 0
\end{aligned}$$

where the last inequality follows from equation 7.1.

$$\begin{aligned}\frac{\partial}{\partial q}\Big[\frac{A(\epsilon,q)}{B(\epsilon,q)}\Big]\Big|_{\epsilon=0,q=1} =& \frac{\frac{\partial}{\partial q}A(\epsilon,q)\cdot B(\epsilon,q) - A(\epsilon,q)\cdot\frac{\partial}{\partial q}B(\epsilon,q)}{B(\epsilon,q)^2}\Big|_{\epsilon=0,q=1} \\ =& \frac{\frac{\partial}{\partial q}A(\epsilon,q)|_{\epsilon=0,q=1}\cdot B(0,1) - A(0,1)\cdot\frac{\partial}{\partial q}B(\epsilon,q)|_{\epsilon=0,q=1}}{B(0,1)^2} \\ =& \frac{A(0,1)}{B(0,1)^2}\frac{\partial}{\partial q}[A(\epsilon,q)-B(\epsilon,q)]|_{\epsilon=0,q=1} > 0\end{aligned}$$

Since $\frac{\partial}{\partial q}A/B$ is a continuous function in ϵ and q, there exists a set $D = \{\epsilon, q : \epsilon \in [0, E], q \in [C, 1]\}$, such that, $\frac{\partial}{\partial q}\left[\frac{A(\epsilon,q)}{B(\epsilon,q)}\right]\Big|_{(\epsilon,q)\in D} > 0$. Therefore, for $E < \epsilon < 1$ and $C < q < 1$, we have $0 < \frac{A(\epsilon,q)}{B(\epsilon,q)} < \frac{A(\epsilon,1)}{B(\epsilon,1)}$. Since $A(\epsilon,1)/B(\epsilon,1) > 1$ (by Theorem 3.3.3), and A/B is continuous in q, we have for $C^* < q < 1$, $A(\epsilon,q)/B(\epsilon,q) > 1$. Hence, for $\max(C, C^*) < q < 1$, $|A(\epsilon,1)/B(\epsilon,1) - 1| > |A(\epsilon,q)/B(\epsilon,q) - 1|$.

7.6 Lemma 7.6.1

Lemma 7.6.1. $\forall m \in \mathbb{R}$ and $\forall a, b \in \mathbb{R}^+$, it holds that

(i) $L_q(ab) = L_q(a) + L_q(b) + (1-q)L_q(a)L_q(b) = L_q(a) + a^{1-q}L_q(b)$.

(ii) $L_q(a^m) = L_q(a)\frac{1-(a^{1-q})^m}{1-a^{1-q}}$.

(iii) $L_q\big(\frac{a}{b}\big) = (\frac{1}{b})^{1-q}(L_q(a) - L_q(b))$.

(iv) $L_q(a)$ is a concave function and $L_q(a) \le a - 1$.

Proof. (i) We know that $L_q(ab) = \frac{(a^{1-q}-1)+(b^{1-q}-1)+(a^{1-q}-1)(b^{1-q}-1)}{1-q}$, which proves (i).

(ii) $L_q(a^m) = \frac{a^{1-q}-1}{1-q}\frac{(a^{1-q})^m-1}{a^{1-q}-1} = L_q(a)\frac{1-(a^{1-q})^m}{1-a^{1-q}}$.

(iii) By (i), we have $L_q(a/b) = L_q(a)/b^{1-q} + L_q(1/b) = [L_q(a) - L_q(b)]/b^{1-q}$.

(iv) We have $\partial^2 L_q(a)/\partial a^2 = -qa^{-q-1} < 0$, hence, $L_q(a)$ is concave. By the mean value theorem of concave function: $L_q(a) - L_q(1) \leq (a-1)\frac{\partial L_q(x)}{\partial x}\Big|_{x=1} \quad \Rightarrow \quad L_q(a) \leq a - 1$. □

7.7 Re-weighting Algorithm for MLqE

The re-weighting algorithm for solving the MLqE in general is described as follows.

To obtain $\hat{\theta}_{\mathrm{ML}q\mathrm{E}}$, we start with an initial estimate $\theta^{(1)}$ which could be any sensible estimate. (We usually use $\hat{\theta}_{\mathrm{MLE}}$ as the starting point.) For each new iteration t $(t > 1)$, $\theta^{(t+1)}$ is computed via

$$\theta^{(t+1)} = \Big\{\theta : 0 = \sum_{i=1}^{n} U(x_i;\theta) f(x_i;\theta^{(t)})^{1-q}\Big\},$$

where $U(x;\theta) = \nabla_\theta \log f(x;\theta) = f'_\theta(x;\theta)/f(x;\theta)$. The algorithm is stopped when a certain convergence criterion is satisfied, for example, the change in $\theta^{(t)}$ is sufficiently small.

To obtain the MLqE for a normal distribution, the above algorithm is simplified

as follows:

$$\hat{\mu}^{(t+1)} = \frac{1}{\sum_{i=1}^{n} w_i^{(t)}} \sum_{i=1}^{n} w_i^{(t)} x_i,$$
$$\hat{\sigma^2}^{(t+1)} = \frac{1}{\sum_{i=1}^{n} w_i^{(t)}} \sum_{i=1}^{n} w_i^{(t)} (x_i - \hat{\mu}^{(t+1)})^2,$$

where $w_i^{(t)} = \varphi(x_i; \hat{\mu}^{(t)}, \hat{\sigma^2}^{(t)})^{1-q}$ and φ is a normal probability density function.

In the M step of the EM-Lq algorithm, the above algorithm is further modified as follows:

$$\mu_j^{(t+1)} = \frac{1}{\sum_{i=1}^{n} \tilde{w}_{ij}^{(t)}} \sum_{i=1}^{n} \tilde{w}_{ij}^{(t)} x_i,$$
$$\sigma_j^{2(t+1)} = \frac{1}{\sum_{i=1}^{n} \tilde{w}_{ij}^{(t)}} \sum_{i=1}^{n} \tilde{w}_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2,$$

where $\tilde{w}_{ij}^{(t)} = \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \varphi(x_i; \mu_j^{(t)}, \sigma_j^{2(t)})^{1-q}$. We iterate the above calculation until $\mu_j^{(t)}$ and $\sigma_j^{2(t)}$ converge, and assign them to μ_j^{new} and $\sigma_j^{2\text{new}}$.

7.8 Proof of Theorem 4.2.3

$$
\begin{aligned}
\sum_{i=1}^{n} L_q(p(x_i;\Psi)) &= \sum_{i=1}^{n} L_q(\sum_{z} p(x_i,z;\Psi)) \\
&= \sum_{i=1}^{n} L_q(\sum_{z} p(z|x_i;\Psi^{\text{old}})\frac{p(x_i,z;\Psi)}{p(z|x_i;\Psi^{\text{old}})}) \qquad (7.5)\\
&\geq \sum_{i=1}^{n}\sum_{z} p(z|x_i;\Psi^{\text{old}}) L_q\Big(\frac{p(x_i,z;\Psi)}{p(z|x_i;\Psi^{\text{old}})}\Big) \qquad (7.6)\\
&= \sum_{i=1}^{n}\sum_{z} p(z|x_i;\Psi^{\text{old}})\frac{L_q(p(x_i,z;\Psi)) - L_q(p(z|x_i;\Psi^{\text{old}}))}{p(z|x_i;\Psi^{\text{old}})^{1-q}} \\
&= \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(p(X,Z;\Psi))}{p(Z|X;\Psi^{\text{old}})^{1-q}} - \frac{L_q(p(Z|X;\Psi^{\text{old}}))}{p(Z|X;\Psi^{\text{old}})^{1-q}}\Big| X = x_i\Big] \\
&= B(\Psi,\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}),
\end{aligned}
$$

where, from equation (7.5) to (7.6), we have used Jensen's inequality on the L_q function due to its concavity (Lemma 7.6.1, part (iv) in Chapter 7). When $\Psi = \Psi^{\text{old}}$, we have

$$
B(\Psi^{\text{old}},\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}}) = A(\Psi^{\text{old}},\Psi^{\text{old}}) = \sum_{i=1}^{n} L_q(p(x_i;\Psi^{\text{old}})).
$$

7.9 Proof of Theorem 4.2.4

Define

$$D(\Psi) = \sum_{i=1}^{n} L_q(p(x_i;\Psi)) - \big(B(\Psi,\Psi^{\text{old}}) + C(\Psi^{\text{old}},\Psi^{\text{old}})\big) \geq 0. \tag{7.7}$$

By theorem 4.2.3, we know that $D(\Psi^{\text{old}}) = 0$ and $D(\Psi) \geq 0$, so $D(\Psi)$ obtains its minimum at $\Psi = \Psi^{\text{old}}$, i.e.,

$$\frac{\partial}{\partial\Psi} D(\Psi)\Big|_{\Psi=\Psi^{\text{old}}} = 0.$$

Taking the derivative of both sides of (7.7), we have the first part of the theorem. Together with equation (4.3) and (4.4), we prove the rest of the theorem.

7.10 Proof of Theorem 4.3.1

For mixture models, we plug equation (4.6) and (4.7) in B,

$$\begin{aligned}
B(\Psi,\Psi^{\text{old}}) &= \sum_{i=1}^{n} E_{\Psi^{\text{old}}}\Big[\frac{L_q(\prod_{j=1}^{k}(\pi_j f_j(X;\theta_j))^{Z_j})}{(\prod_{j=1}^{k}(\frac{\pi_j^{\text{old}} f_j(X;\theta_j^{\text{old}})}{f(X;\Psi^{\text{old}})})^{Z_j})^{1-q}}\Big| X = x_i\Big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} L_q(\pi_j f_j(x_i;\theta_j)) \cdot \frac{p(Z_j = 1, Z_{-j} = 0 | X = x_i;\Psi^{\text{old}})}{(\frac{\pi_j^{\text{old}} f_j(x_i;\theta_j^{\text{old}})}{f(x_i;\Psi^{\text{old}})})^{1-q}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_j(x_i,\Psi^{\text{old}})^q L_q(\pi_j f_j(x_i;\theta_j)).
\end{aligned}$$

7.11 Proof of Theorem 4.3.2

Apply the first order condition on B (note $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$),

$$\frac{\partial}{\partial \theta_j} B(\Psi, \Psi^{\text{old}}) = \sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \frac{\frac{\partial}{\partial \theta_j} f_j(x_i; \theta_j)}{f_j(x_i; \theta_j)} (\pi_j f_j(x_i; \theta_j))^{1-q}, \tag{7.8}$$

$$\begin{aligned} \frac{\partial}{\partial \pi_j} B(\Psi, \Psi^{\text{old}}) &= \sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \frac{f_j(x_i; \theta_j)}{(\pi_j f_j(x_i; \theta_j))^q} - \sum_{i=1}^{n} \tilde{\tau}_k(x_i, \Psi^{\text{old}}) \frac{f_k(x_i; \theta_k)}{(\pi_k f_k(x_k; \theta_k))^q} \\ &= \sum_{i=1}^{n} \frac{\tilde{\tau}_j(x_i, \Psi^{\text{old}}) f_j(x_i; \theta_j)^{1-q}}{\pi_j^q} - \sum_{i=1}^{n} \frac{\tilde{\tau}_k(x_i, \Psi^{\text{old}}) f_k(x_i; \theta_k)^{1-q}}{\pi_k^q}, \end{aligned} \tag{7.9}$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) \frac{\frac{\partial}{\partial \theta_j} f_j(x_i; \theta_j)}{f_j(x_i; \theta_j)} f_j(x_i; \theta_j)^{1-q} \quad \text{and} \tag{7.10}$$

$$\pi_j \propto \Big[\sum_{i=1}^{n} \tilde{\tau}_j(x_i, \Psi^{\text{old}}) f_j(x_i; \theta_j)^{1-q} \Big]^{\frac{1}{q}}.$$

7.12 Proof of Theorem 4.2.4 for the mixture model case

By equations (7.8) and (7.9), the derivatives of B at $\Psi = \Psi^{\text{old}}$ are

$$\begin{aligned}
&\frac{\partial}{\partial \theta_j} B(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} \\
&= \sum_{i=1}^{n} \Big(\frac{\pi_j^{\text{old}} f_j(x_i;\theta_j^{\text{old}})}{f(x_i;\Psi^{\text{old}})}\Big)^q \cdot \frac{\frac{\partial}{\partial\theta_j} f_j(x_i;\theta_j)\big|_{\theta_j=\theta_j^{\text{old}}}}{f_j(x_i;\theta_j^{\text{old}})} (\pi_j^{\text{old}} f_j(x_i;\theta_j^{\text{old}}))^{1-q} \\
&= \sum_{i=1}^{n} \frac{\pi_j^{\text{old}} \big(\frac{\partial}{\partial\theta_j} f_j(x_i;\theta_j)\big|_{\theta_j=\theta_j^{\text{old}}}\big)}{f(x_i;\Psi^{\text{old}})^q}, \\
&\frac{\partial}{\partial \pi_j} B(\Psi, \Psi^{\text{old}})\Big|_{\Psi=\Psi^{\text{old}}} \\
&= \sum_{i=1}^{n} \Big(\frac{\pi_j^{\text{old}} f_j(x_i;\theta_j^{\text{old}})}{f(x_i;\Psi^{\text{old}})}\Big)^q \frac{f_j(x_i;\theta_j^{\text{old}})^{1-q}}{(\pi_j^{\text{old}})^q} - \sum_{i=1}^{n} \Big(\frac{\pi_k^{\text{old}} f_k(x_i;\theta_k^{\text{old}})}{f(x_i;\Psi^{\text{old}})}\Big)^q \frac{f_k(x_i;\theta_k^{\text{old}})^{1-q}}{(\pi_k^{\text{old}})^q} \\
&= \sum_{i=1}^{n} \frac{f_j(x_i;\theta_j^{\text{old}}) - f_k(x_i;\theta_k^{\text{old}})}{f(x_i;\Psi^{\text{old}})^q}.
\end{aligned}$$

We calculate the derivatives of $\sum_{i=1}^{n} L_q(p(x_i;\Psi))$ at $\Psi = \Psi^{\text{old}}$,

$$\begin{aligned}
\frac{\partial}{\partial\theta_j} \sum_{i=1}^{n} L_q(p(x_i;\Psi))\Big|_{\Psi=\Psi^{\text{old}}} &= \sum_{i=1}^{n} \frac{\pi_j^{\text{old}} \big(\frac{\partial}{\partial\theta_j} f_j(x_i;\theta_j)\big|_{\theta_j=\theta_j^{\text{old}}}\big)}{f(x_i;\Psi^{\text{old}})^q}, \\
\frac{\partial}{\partial\pi_j} \sum_{i=1}^{n} L_q(p(x_i;\Psi))\Big|_{\Psi=\Psi^{\text{old}}} &= \sum_{i=1}^{n} \frac{f_j(x_i;\theta_j^{\text{old}}) - f_k(x_i;\theta_k^{\text{old}})}{f(x_i;\Psi^{\text{old}})^q}.
\end{aligned}$$

By comparing the formulas above, we obtain the first equation of the theorem. Together with equation (4.3) and (4.4), we prove the rest of the theorem.

Bibliography

[1] D. Ferrari and Y. Yang, "Maximum Lq-likelihood estimation," *Annals of Statistics*, vol. 38, pp. 753–783, 2010.

[2] P. J. Huber, "A robust version of the probability ratio test," *Annals of Mathematical Statistics*, vol. 36(6), pp. 1753–1758, 1965.

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, 1st ed. Wiley, 1986.

[4] P. J. Huber and E. M. Ronchetti, *Robust statistics*, 2nd ed. Wiley, 2009.

[5] P. J. Bickel and K. A. Doksum, *Mathematical Statistics Basic Ideas and Selected Topics Volume I*, 2nd ed. Pearson Prentice Hall, 2007.

[6] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50 (1), pp. 1–25, 1982.

[7] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and

the em algorithm," *Society for Industrial and Applied Mathematics Review*, vol. 26 (2), pp. 195–239, 1984.

[8] R. A. S. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London Series A*, vol. 222, pp. 309–368, 1922.

[9] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1(6), pp. 80–86, 1945.

[10] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18(1), pp. 50–60, 1947.

[11] J. Arbuthnot, "An argument for divine providence, taken from the constant regularity observed in the births of both sexes," *Philosophical Transactions of the Royal Society of London*, vol. 27(325-336), pp. 186–190, 1710.

[12] E. L. Lehmann and H. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*, 1st ed. Springer, 2006.

[13] A. R. Cushny and A. R. Peebles, "The action of optical isomers ii. hyoscines," *Journal of Physiology*, vol. 32(5-6), pp. 501–510, 1905.

[14] R. G. Staudte and S. J. Sheather, *Robust Estimation and Testing*, 1st ed. Wiley, 1990.

[15] R. Beran, "Minimum hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5(3), pp. 445–463, 1977.

[16] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 1–20, 2000.

[17] C. F. J. Wu, "On the convergence properties of the em algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[18] M. P. Windham and A. Cutler, "Information ratios for validating mixture analyses," *Journal of the American Statistical Association*, vol. 87, pp. 1188–1192, 1992.

[19] W. Guo and S. Cui, "A q-parameterized deterministic annealing em algorithm based on nonextensive statistical mechanics," *IEEE Transactions on Signal Processing*, vol. 56 (7), pp. 3069–3080, 2008.

[20] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[21] L. A. Garcia-Escudero and A. Gordaliza, "Robustness properties of k-means and trimmed k-means," *Journal of the American Statistical Association*, vol. 94, pp. 956–969, 1999.

[22] Y. Qin and C. E. Priebe, "Maximum lq-likelihood estimation via the expectation

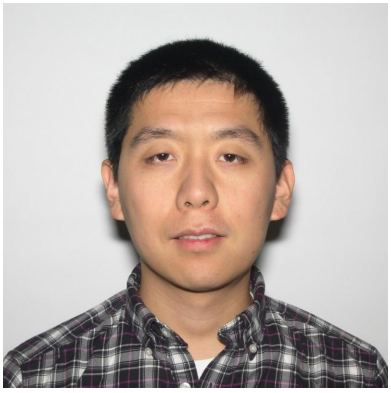
maximization algorithm: A robust estimation of mixture models," *Journal of the American Statistical Association*, vol. 108(503), pp. 914–928, 2013.

[23] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe, "A consistent adjacency spectral embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, vol. 107(499), pp. 1119–1128, 2012.

[24] W. R. Gray, J. A. Bogovic, J. T. Vogelstein, B. A. Landman, J. L. Prince, and R. J. Vogelstein, "Magnetic resonance connectome automated pipeline," *IEEE Pulse*, vol. 3(2), pp. 42–48, 2012.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, pp. 1–38, 1977.

Vita



Yichen Qin received a BS degree in Statistics from Renmin University of China in 2005, an MA degree in Statistics from Columbia University in 2007, and enrolled in the Applied Mathematics and Statistics Ph.D. program at Johns Hopkins University in 2008. His research focuses on robust statistics, computational statistics, mixture models and EM algorithms.

Beginning August 2013, Yichen will serve as Assistant Professor at the University of Cincinnati in Ohio.